The Construction and Validation of Rating Scales for Oral Tests in English as a Foreign Language

by

Glenn Fulcher

B.D., M.Th., A.K.C. (University of London)
M.A. in Applied Linguistics (University of Birmingham)
P.G.C.E. (University of Cambridge)

Submitted for the degree of Ph.D. in the Department of Linguistics and Modern English Language, University of Lancaster.

January 1993
The Construction and Validation of Rating Scales for Oral Tests in English as a Foreign Language

by

Glenn Fulcher

B.D., M.Th., A.K.C. (University of London)
M.A. in Applied Linguistics (University of Birmingham)
P.G.C.E. (University of Cambridge)

Abstract

The present research investigates the principles upon which rating scales in oral testing are constructed and used, and the subsequent claims of reliability and validity made for them. The research addresses two main questions: (i) can rating scales be constructed on the basis of an analysis of a database of actual student speech; and (ii) are scales produced on the basis of student speech superior to those produced using a-priori methods?

A corpus of spoken data was collected and analyzed. Discriminant analysis was used in order to isolate factors which could discriminate between students of different ability levels, and a Fluency and an Accuracy rating scale constructed. The Fluency rating scale was seen to be the most stable in the construction phase of the study.

Forty seven students took three tasks. Video recordings were rated by five raters on the two rating scales and the English Language Testing Service rating scale.
Use of questionnaires and retrospective reports indicated that affective facets did not contaminate results.

The reliability of the three rating scales was assessed using a G-study, traditional correlational methods, and a Rasch Partial Credit model. It was discovered that the Fluency rating scale was used most reliably.

Validity was investigated in two concurrent studies, a group differences study, a Multitrait-Multimethod study with Confirmatory Factor Analysis, and a Rasch Partial Credit study. Divergent validity could not be established for any of the rating scales. However, a new Rasch validity statistic which can be related to bands rather than entire scales was developed, and results of its application showed that the Fluency rating scale possessed both coherence and continuum validity in three bands.

Conclusions suggest that databased approaches to rating scale development are promising, and suggestions are made for future research into band and rating scale construction.
Acknowledgements

Research into oral testing cannot use vast numbers of students and teachers unless there are no restrictions on time and funding. The forty seven students used in this study gave up a considerable amount of time, and although they cannot be named I am grateful to them for their willingness to take part. Each of the five raters who volunteered to view all the video cassettes and assign three grades to each student on each task spent an estimated 40 hours plus 30 minutes in the debriefing interview and another 10 hours to complete the second marking for the intra-rater study: a daunting amount of work for busy professionals. Without the help of these five teachers, who shall remain anonymous throughout the text of this study, it would have been impossible to complete a project on oral testing.

Thanks are due to Peter Hargreaves of the University of Cambridge Local Examinations Syndicate for providing the data upon which the concurrent study with First Certificate results was based, and Mike Milanovic and Nick Saville for comments on the first draft of that study. I should also like to acknowledge the cooperation of Nick Butler of the British Council and Clive Mogford (then Director of the British Council in Cyprus) for giving permission and providing facilities to make the recordings of English Language Testing Service interviews which formed the database for the studies in Chapters 4 and 5.

-iv-
I would like to thank Brian Francis of the Department of Applied Statistics at Lancaster University who, despite constantly being badgered by research students, always found time to discuss any problems I came across in the course of my work. In particular, I would like to thank him for his comments on the form and implications of the Rasch Partial Credit validity statistic for bands developed in Chapter Nine.

Working away from Lancaster with only brief summer visits often led to a feeling of isolation. The lack of people with whom problems encountered can be discussed face-to-face can be depressing for students working at a distance, and this may lead to a failure to complete. That this study has been completed is due entirely to my supervisor, Charles Alderson. A letter at just the right moment, critical but encouraging comments on drafts, plans and data, and suggestions which have widened my horizons and ability to conceptualise within the field of language testing. I could not have wished for a better supervisor, and I will always be deeply indebted and grateful to him.

Finally, I would like to thank my family. Firstly, my mother and father who volunteered to pay all costs incurred, and never complained at the loss of their car whenever I visited Lancaster! They have provided the best support I could possibly have asked for, both in terms of encouragement and financial help. My wife, Marie, has tolerated a great deal of neglect during the last few
years in order for me to complete this study. Not only has she been supportive and understanding, but also undertook the compilation of the bibliography and helped with proofreading. Lastly, Greg, who did not demand so much of my attention during the first few months of life that I could not settle down to revise the final drafts.
Table of Contents

Chapter One
Introduction 1
Section 1: Issues in oral language testing 1
Section 2: Objectives of this research 4
Section 3: Organisation of the thesis 7

Chapter Two
Issues in Oral Testing 8
Section 1: Issues from the history of oral testing 10
  1.1. The Foreign Service Institute: the beginning of oral testing in the United States Government in the post-war period 10
  1.2 The development of oral testing outside the United States Government 13
  1.3 FSI to ACTFL: The development of oral rating scales 16
    1.3.1 The FSI rating scale 16
    1.3.2 Critique of the FSI rating scale 23
    1.3.3 The ILR rating scale 26
    1.3.4 The ACTFL/ETS rating scale 27
  1.4 Discussion 32
Section 2: Issues in reliability and validity in oral testing 41
  2.1 Reliability in oral testing 41
    2.1.1 Evidence from the United States 41
    2.1.2 Evidence from other countries 47
    2.1.3 Reliability and rater training 49
    2.1.4 A methodological note on reliability studies: some recent developments 50
    2.1.5 Conclusions 51
  2.2 Validity in oral testing 52
    2.2.1 Validity claims and some empirical evidence 54
    2.2.2 Concurrent validation 63
    2.2.3 Controversy over band descriptors in oral rating scales 76
    2.2.4 Conclusions 89
Section 3: Rating scale construction in oral testing 91
  3.1 Approaches to constructing rating scales 91
  3.2 Psychometric considerations 100
  3.3 Terminology used in constructing band descriptors 101
  3.4 A special case of terminological confusion in oral rating scales: the Native Speaker Yardstick 107
Chapter Seven
Student Affective Factors 247
Section 1: Introduction 247
Section 2: Questionnaire structure 250
Section 3: Summary statistics for the questionnaires 252
Section 4: Analysis of questionnaire responses by group and task preference 254
Section 5: Student self assessment 259
Section 6: The relationship of reported anxiety to Rasch ability estimates 260
Section 7: Summary of qualitative analysis of responses to open ended questions 260
Section 8: Discussion 263

Chapter Eight
Reliability 266
Section 1: Introduction 266
Section 2: Generalizability study 268
Section 3: Correlational analysis 273
  3.1 Reliability coefficients on scales 274
  3.2 Reliability coefficients across tasks 276
Section 4: Rasch Partial Credit analysis 277
  4.1 Introduction 277
  4.2 Rating scales and rater bias 278
  4.3 A Partial Credit model to analyze rating patterns 282
  4.4 The effect of task on rating patterns 287
  4.5 Task difficulty 292
  4.6 Partial Credit reliability estimates 292
Section 5: Intra-rater reliability 293
Section 6: Discussion 294
  6.1 Task 3 294
  6.2 Inter- and intra-rater reliability 297
  6.3 Implications for rater training 298

Chapter Nine
Validity 302
Section 1: Introduction 302
Section 2: Two concurrent studies 302
  2.1 Introduction to study one 302
  2.2 Study one: FCE - Developmental tests concurrent study 303
    2.2.1 The five FCE papers, Fluency, Accuracy and ELTS 303
    2.2.2 FCE oral rating components, Fluency, Accuracy and ELTS 306
    2.2.3 Discussion 308
  2.3 Study two: The relationship between teacher assessments of student oral abilities and the developmental scales 310
2.3.1 Introduction 310
2.3.2 Available data from teacher assessments 313
2.3.3 Analysis of concurrent teacher assessments 314

Section 3: Validity by Group Differences 315
Section 4: Construct Validity 317
4.1 A Multitrait - Multimethod study 317
4.2 Maximum Likelihood Confirmatory Factor Analysis 320
4.3 Rasch Partial Credit analysis 324
4.4 Further validity considerations 332
4.5 Discussion and conclusions 334

Chapter Ten
Retrospective Report Study 339
Section 1: Introduction 339
Section 2: Interviews with raters 341
Section 3: Interviews with students 343
Section 4: Analysis of data 344
4.1 Rater responses 344
4.2 Student responses 359
4.2.1 Ability level of the sample 359
4.2.2 Summary of findings 359

Chapter Eleven
Discussion and Conclusions 361
Section 1: Preliminaries 361
Section 2: Validity and the use of the data-driven approach to scale development 362
2.1 The lack of divergent validity 362
2.2 Validity in the Partial Credit model study 366
2.2.1 A methodological note 366
2.2.2 Coherence and Continuum validity 368
2.2.3 Conclusions 371
Section 3: Assumptions of this research 373
Section 4: Suggestions for future research 376
4.1 Validation of component traits 376
4.2 Task and topic 379
4.3 Specificity and length of rating scales 381
4.4 Validation methodology and tools 382
4.5 Research into other rating criteria 383
Section 5: Conclusions 384

References 385
Appendix to Chapter Two 412
Appendix to Chapter Four 415
Appendix to Chapter Five 424
Appendix to Chapter Six 435
Appendix to Chapter Seven 446
Appendix to Chapter Eight 460
Appendix to Chapter Nine 472
Appendix to Chapter Ten 481

List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig. 2.1</td>
<td>FSI rating factors</td>
<td>17</td>
</tr>
<tr>
<td>Fig. 2.2</td>
<td>FSI weighting table</td>
<td>18</td>
</tr>
<tr>
<td>Fig. 2.3</td>
<td>FSI conversion table</td>
<td>18</td>
</tr>
<tr>
<td>Fig. 2.4</td>
<td>A comparison of the ILR and ACTFL rating scales</td>
<td>28</td>
</tr>
<tr>
<td>Fig. 4.1</td>
<td>The relationship between actual band scores and predicted band scores for the explanatory fluency categories</td>
<td>152</td>
</tr>
<tr>
<td>Fig. 4.2</td>
<td>Category 1 means</td>
<td>154</td>
</tr>
<tr>
<td>Fig. 4.3</td>
<td>Category 2 means</td>
<td>154</td>
</tr>
<tr>
<td>Fig. 4.4</td>
<td>Category 3 means</td>
<td>155</td>
</tr>
<tr>
<td>Fig. 4.5</td>
<td>Category 4 means</td>
<td>155</td>
</tr>
<tr>
<td>Fig. 4.6</td>
<td>Category 5 means</td>
<td>156</td>
</tr>
<tr>
<td>Fig. 4.7</td>
<td>Category 6 means</td>
<td>156</td>
</tr>
<tr>
<td>Fig. 4.8</td>
<td>Category 7 means</td>
<td>157</td>
</tr>
<tr>
<td>Fig. 4.9</td>
<td>Category 8 means</td>
<td>157</td>
</tr>
<tr>
<td>Fig. 5.1</td>
<td>The relationship between actual band scores and predicted band scores for accuracy categories</td>
<td>203</td>
</tr>
<tr>
<td>Fig. 5.2</td>
<td>Category 1 means</td>
<td>209</td>
</tr>
<tr>
<td>Fig. 5.3</td>
<td>Category 2 means</td>
<td>209</td>
</tr>
<tr>
<td>Fig. 5.4</td>
<td>Category 3 means</td>
<td>210</td>
</tr>
<tr>
<td>Fig. 5.5</td>
<td>Category 4 means</td>
<td>210</td>
</tr>
<tr>
<td>Fig. 5.6</td>
<td>Category 5 means</td>
<td>211</td>
</tr>
<tr>
<td>Fig. 5.7</td>
<td>Category 6 means</td>
<td>211</td>
</tr>
<tr>
<td>Fig. 5.8</td>
<td>Category 7 means</td>
<td>212</td>
</tr>
<tr>
<td>Fig. 5.9</td>
<td>Category 8 means</td>
<td>212</td>
</tr>
<tr>
<td>Fig. 5.10</td>
<td>Category 9 means</td>
<td>213</td>
</tr>
<tr>
<td>Fig. 5.11</td>
<td>The suggested natural order of acquisition from the morpheme studies</td>
<td>220</td>
</tr>
</tbody>
</table>
Fig. 5.12 A possible acquisition hierarchy for eight aspects of grammar
Fig. 5.13 The relationship between acquisition groups identified in fig. 5.12 and the band descriptors in the Accuracy rating scale
Fig. 8.1 Rater agreement (Fluency)
Fig. 8.2 Task comparability (Fluency)
Fig. 8.3 Rater agreement (Accuracy)
Fig. 8.4 Task comparability (Accuracy)
Fig. 8.5 Rater agreement (ELTS global assessment)
Fig. 8.6 Task comparability (ELTS global assessment)
Fig. 8.7 Rank order of judge severity on all rating scales
Fig. 8.8 Rating scale structure
Fig. 8.9 Item Characteristic Curves for five raters on the Fluency rating scale
Fig. 8.10 Item Characteristic Curves for five raters on the Accuracy rating scale
Fig. 8.11 Item Characteristic Curves for five raters on the ELTS global rating scale
Fig. 8.12 Item Characteristic Curves for five raters on the Fluency rating scale, taking task variability into account
Fig. 8.13 Item Characteristic Curves for five raters on the Accuracy rating scale, taking task variability into account
Fig. 8.14 Item Characteristic Curves for five raters on the ELTS global rating scale, taking task variability into account
Fig. 8.15 Reliability statistics for each rating scale by raters and tasks
Fig. 9.1 Path diagram for an ML solution of three intercorrelated test method facets
Fig. 9.2 Item Characteristic Curves for the study of rating scale validity
Fig. 9.3 A Rasch Partial Credit validity curve for use with bands in rating scales
Chapter One

Introduction

Section 1: Issues in oral language testing

Oral language testing has been an important area of research and development in language testing since the Second World War, but most of the published research on issues in oral testing has only appeared in the last fifteen years. One reason for this may be that language testers had generally assumed that oral tests could not meet the psychometric criteria for test development in general (Lado, 1961). However, it is now widely recognised that it is important to test oral ability given its growing importance in language syllabuses, and new methods and techniques for the assessment of the psychometric properties of oral tests have been developed (see, for example, Masters, 1990).

The issues surrounding oral testing are complex and often controversial. An oral test, like any test, essentially comprises an elicitation technique and a scoring procedure (Alderson, 1983: 91). We wish to define these two categories with slightly more precision. We make a distinction between elicitation technique which refers to the specific prompts, questions or comments which the interviewer or controller uses to elicit a language sample, and a "task" which refers to the format in which the elicitation techniques are employed. Further, there is the issue of "topic", which refers to the content or subject of the task. It seems possible
that the choice of topic may have an effect on student performance on any given task using any specific elicitation techniques. The elicitation technique, the task and the topic together constitute the "situation" into which the student is placed for the purpose of oral testing, and may influence the type of language which students produce, or the expected discourse outcomes.

When discussing scoring procedures we are concerned with the way in which the scoring system, usually a rating scale, is developed and used. Two related issues here are the criteria selected against which the students are to be rated and, given these criteria the number of bands or categories in the rating scale which can be justified. Needless to say, there is then the knotty problem of how the rating criteria are to be related to the language sample obtained from the test: are we to develop so-called "common measures" which can relate to any language and any testing situation, or "specific measures" which relate to only one testing situation or, in an English for Specific Purposes (ESP) context, to one type of predicted outcome in terms of the language sample. This raises the issue of the generalizability of results to situations outside that of the test itself. Further, is it preferable to develop oral rating scales which are global/holistic or componential/analytic, and what evidence can be supplied which suggests that one method is preferable to the other given the extent of our current knowledge about what makes up the construct "oral
The reliability, validity and practicality of any oral test are clearly of major concern, and relate to all aspects of the issues outlined above. The training of raters is one area of reliability which will be discussed to some extent in this study.

Another area of concern is that of test method facets which may confound studies of reliability and validity because of their effect on scores. Test method facets include not only the task, elicitation technique and the topic, but affective factors of which test anxiety has often been hypothesized to be the most important, as well as raters and their degree of training in the use of a rating scale.

Other issues which are dealt with in the oral testing literature include the status of the interview as "real conversation" in "real-life" interaction, as opposed to a test "genre" in which we only observe "test-type language", the degree to which rating scales distinguish between "performance" (as in many so-called "direct tests") and "competence" (are they "indirect" in that the purpose is to measure underlying ability?), and the washback effect which any testing procedure may have on a teaching curriculum.

All of these issues are worthy of research. However, it is clearly not possible to deal with everything we do not know about the process of oral testing in one study. This study will concern itself with one area of oral
testing, namely the construction and validation of oral rating scales. Validation procedures necessarily include the provision of as much evidence as possible to indicate whether a rating procedure is or is not valid (Messick, 1989a), and so although some of the issues outlined above are not of primary focus, they will be dealt with to some extent in the body of the thesis.

Section 2: Objectives of this research

Research into the construction and operation of oral rating scales was suggested by findings which indicate that there is frequently little connection between the notion of the linguistic/communicative development of the non-native speaker underlying band descriptors in rating scales and what real speakers say in real situations, be these informal conversations (Fulcher, 1987) or tests (Fulcher, 1989). This study sets out to investigate (i) whether scales can be constructed on the basis of an analysis of a database of actual student speech; and (ii) whether scales produced on the basis of a database of student speech are superior to those produced using a-priori methods.

Problems created in the production of rating scales on an a-priori basis are investigated at some length. The tradition of oral testing in the United States was taken as being typical of current a-priori approaches to rating scale development, and these approaches have been copied in other parts of the world. American tests are therefore
scrutinised in some detail, mainly in Chapter 2, although Australian and European tests are considered when available published research is relevant to scale construction.

The initial step in the research reported in this study was to construct a database of student speech and analyze it for evidence of factors related to the constructs of Fluency and Accuracy which may have influenced the band score achieved by the student. Discriminant analysis was used to predict student scores and differences between the scores of students of different ability levels. It was discovered that only very high prediction rates from factors related to performance to test scores could legitimize the use of this procedure in the construction of rating scales.

The influence of test method facets on test scores and their interpretation is now recognised (Bachman, 1990). In this project great care was taken either to control or account for any test method facets which could have rendered any conclusions useless because of contamination of data. Similarly, high reliability statistics are reported for all the rating scales being used. Had this not been the case, it would not have been legitimate to proceed to the investigation of the validity of the rating scales.

In the validation study, little evidence could be found of concurrent or discriminant validity for either the Fluency or the Accuracy rating scale. However, the
Rasch Partial Credit study did indicate that the Fluency scale in particular was being used in a much more stable way than either the Accuracy or the English Language Testing Service rating scale. It was suspected that the methods used in the validation of the rating scale were not sufficiently sensitive to isolate observed consistencies across tasks and raters, and a new validity statistic was developed which allows the test researcher to assign a validity statistic not to a scale as a whole, but to each band within a scale. The concepts of coherence and continuum validity are outlined as a theoretical counterpart to this new statistic.

It was discovered that three of the bands in the Fluency rating scale are behaving in the way which one would normally expect from a valid measurement instrument, but that the other bands were not. These consistencies can be directly related to the excellent prediction pattern observed in the discriminant analysis used in the construction of the rating scale. That is, the high observed validity coefficients are determined by the success of the database analysis and construction process. The approach to validating bands developed in this research also allows the test constructor to suggest why a band or a scale is not working in the way expected when first developed.

The conclusions of this research suggest that, although the databased approach to rating scale construction is much more time consuming than traditional
a-priori methods, it has many potential advantages. The most important of these is that results from validation studies can be related directly to claims about the coherence of any second language acquisition model which underlies the rating scale. Further, individual bands may be assessed and their place in the model evaluated by reference to the descriptors used.

It is argued that the approach developed here is significant in its findings, and could lead to future research which would extend our knowledge of scale and band construction to other components of the construct "oral ability." The implications for oral testing and suggestions for future research are discussed.

Section 3: Organisation of the thesis

All figures are included in the text of the thesis for immediate reference. Tables are reserved for the Appendices. This policy was adopted in order to remove multiple tables from the text, which may have been irritating for the reader.

Each Appendix bears the number of the Chapter to which it is attached. Thus, for example, there is no Appendix One, as no additional material is required for Chapter One. The first Appendix is the Appendix to Chapter Two.
Chapter Two

Issues in Oral Testing

The importance of oral testing has been recognised within the language teaching profession since the Second World War, when it was realised that many American service personnel and diplomats did not have the linguistic skills necessary to carry out their duties effectively. Since that War many tests of oral abilities have been developed, firstly in the United States and then in Europe and elsewhere, generally in the case of Europe as a part of a larger test battery. It is because oral testing has a longer history in the United States, because researchers in the United States have generated more published studies, and because many other widely used tests are based upon American models, that we shall concentrate on rating scales developed in the United States.

The emphasis, from the very beginning, was placed upon developing oral tests which would be capable of providing information on the candidate which could be used to predict his/her success in some future real-life situation, as this quotation from Kaulfers (1944: 137, quoted in Barnwell, 1987: 35) makes clear:

"The nature of the individual test items should be such as to provide specific, recognisable evidence of the examinee’s readiness to perform in a life-situation, where lack of ability to understand and speak extemporaneously might be
a serious handicap to safety and comfort, or to the effective execution of military responsibilities."

This general aim has not really changed (Skehan, 1988: 211), although the methodology and techniques used in oral testing have developed considerably. Much of the work which has been done in the field of oral testing in recent years has undoubtedly been influenced by developments which took place in the United States, and many of the issues which are now attracting considerable attention relate directly to the new generation of oral tests which stem from the early post-war period. It is also true to say that many organisations developing oral tests outside the United States used the Foreign Service Institute (FSI) test and its descendants as a model upon which to build. This is true, for example, of the Australian Second Language Proficiency Ratings (Ingram and Wylie, 1985).

This chapter will therefore begin by considering the history of oral testing in the United States in order to highlight some of the issues which are still of great concern in oral test and rating scale development (Section 1). Secondly, issues in the study of reliability and validity will be considered (Section 2). Thirdly, the way in which rating scales for oral tests are constructed will be reviewed and assessed (Section 3). It will be argued throughout that although the issues to be outlined in oral testing have been discussed for a considerable
length of time there is a lack of empirical evidence to support coherent oral testing theory or practice, especially with regard to rating scales. It will also be argued that in the field of oral testing there has been a trend for writers and test developers to defend their views and systems by claiming high face validity.

Section 1. Issues from the History of Oral Testing.

1.1 The Foreign Service Institute; the beginning of oral testing in the United States Government in the post-war period.

Prior to the Second World War the American Council of Learned Societies ran language programmes for military personnel and diplomats in "languages of potential military and diplomatic importance" (Liskin-Gasparro, 1984b: 17), but with the realisation that American personnel were ill-equipped to deal with the linguistic demands which were placed upon them during the war, the FSI was set up in order to teach foreign languages which would be needed in overseas posts (Kramsch, 1986b: 366).

The initial impetus to the development of oral tests was undoubtedly military need, but it was not until 1952 that it was realised that it would be necessary to measure the oral skills taught by the FSI Instructors when the Civil Service Commission decided to create a register of personnel documenting their familiarity with foreign languages and cultures (Liskin-Gasparro, op. cit., 18 - 19). A committee with the responsibility for developing a test of oral abilities produced the first
oral rating scale which consisted of six bands. Band 1 represented no ability to use a language, while band 6 represented native speaker ability. No information is available concerning the choice of 6 bands as opposed to any other number, and no evidence is available to explain why a minimum of band 4 was required for diplomatic personnel. Sollenberger (1978: 5) states that "to avoid complicating the task, no effort was made to separate the components of language proficiency...." What was developed was thus an intuitive six band holistic rating scale with "weak" descriptors only for the lowest and highest bands. The number of bands to be included in a rating scale, whether it should be holistic or componential, and how the descriptors are to be generated are all major current areas of concern.

Political and bureaucratic problems held up work on and implementation of the 1952 testing procedures until 1956, when a survey discovered that personnel still did not possess the required oral skills for their roles, and in 1956 the FSI was given the responsibility to provide evidence of foreign language proficiency for all foreign service personnel. Sollenberger (1978) reports that one of the initial problems which the FSI faced was that the rank and age of officers tended to influence the raters' judgements; the FSI developers had identified one aspect of what today would be called test bias.

In 1958 the FSI testing unit further developed the 1952 scales by adding a checklist of five factors for
oral raters, each measured on the six point scale. These five factors were accent, comprehension, fluency, grammar and vocabulary (Adams, 1980). This was the first step towards developing componential rating, even though the components were to be used as a check on a single holistic mark. Further, although it was claimed that the rating procedure was "a highly accurate predictor" the limitation of the system was also acknowledged to be that it did not measure "effective communication" (Sollenberger, 1978: 7 - 8). Thus, from the very earliest days in the development of modern oral rating scales the roles of linguistic competence and communicative ability were issues of which the testers were aware.

Confidence in the new oral testing procedures developed by the FSI was so high that during the 1960s the method was adopted (and adapted) by the Defense Language Institute, the Central Intelligence Agency (CIA) and the Peace Corps. Quinones (no date) details the way in which the CIA adapted the FSI system for its own use, the most important feature of which was the use of multiple raters and an averaging system for their marks in an attempt to increase reliability. In 1968 these diverse agencies came together to produce a standardized version of the levels, which today is known as the Interagency Language Roundtable (ILR) (Lowe, 1987).

These early developments in oral testing thus generated interest in holistic versus componential rating, bands and their descriptors, test bias, the
difference between linguistic and communicative criteria for rating, and reliability. All of these areas are very high on the agenda of researchers today.

1.2 The development of oral testing outside the United States Government.

The substantial use of the FSI system by the Peace Corps outside the control of the Government agencies introduced the new oral testing system to academics and teachers (Liskin-Gasparro, 1984b: 22), and in the 1970s it was adopted by many universities and states for the purpose of bilingual teacher certification. The process of introduction of the FSI system into universities and schools began with the development of the Testing Kit Workshops (Adams and Frith, 1979), initially for teachers of French and Spanish. Inter-rater reliabilities for oral scores between teachers and FSI raters were reported to be consistently over .84. The teachers used in this rating process had not been subjected to a lengthy period of rater training, and the confidence shown in the FSI system most probably came, in part, through the assumption that the rating scales in themselves possessed a degree of psychological reality rather than being a mere artefact of the training. The importance of using untrained raters in validity studies is highlighted by Alderson (1991b), who uses the word "cloning" to describe rater training. Once training has taken place it could be argued that test results, upon which quantitative
research into validity is conducted, is prejudiced. That is, the training of raters to use a rating scale prior to its validation ensures that raters use the scale in the same way, irrespective of whether the descriptors in the bands actually reflect or describe the trait which the rating scale was designed to measure by its constructors. This initial situation contrasts with the present approach of the American Council on the Teaching of Foreign Languages (ACTFL) which consistently states that the system can only be operated after a full and lengthy period of training in the use of the ACTFL scales (Barnwell, 1989). We will return to the issue of rater training in Section 2.1.3 on page 49.

Barnwell (1987: 36) provides three other reasons why the FSI approach to oral testing became generally popular within a short space of time. Firstly, as a direct test of oral ability, the FSI was seen to have high face validity. Unfortunately, claims of high face validity have often resulted in claims of overall validity for oral tests since, with face validity often being projected as the only type of validity which needs to be seriously considered (Underhill, 1987). Others have warned of the dangers of this approach (Stevenson, 1981; Messick, 1989a). Secondly, inter-rater reliability had been shown to be high. Thirdly, by the late 1970s there was a growing interest in the notional/functional approach to teaching English, and it was widely perceived that the direct oral test was a natural testing adjunct
of new teaching methods. Thus, the relationship between teaching methodology and testing was recognised, and the importance attached to face validity may be seen in that context.

In 1979 the work of J. B. Carroll (1967) became the focus of considerable attention. Carroll had administered the oral test to college majors of French, German, Russian and Spanish in the United States, and concluded that very few college majors in these foreign languages were capable of achieving a level above 2/2+, level 3 in the FSI rating system being the minimum for professional working proficiency. Carroll’s study was replicated in 1979 by the Educational Testing Service (ETS) (Liskin-Gasparro, 1984b: 27), and it was argued that if the ILR approach to oral testing were to be used by universities, colleges and schools, the rating scales would have to be developed to allow more discrimination below the ILR 2+ level. The rationale was simply that it would not be appropriate for students to spend many hours of study on the foreign language and register no progress at all on the rating scale. This led directly to the involvement of the ETS and the growth of the importance of the ACTFL to alter the ILR rating scales to suit the new purposes to which oral tests were to be put outside the government (Liskin-Gasparro, 1984a: 447; 1984b: 25; Lowe, 1983: 232; Lowe, 1985a: 13; Lowe, 1987; Clark 1988a: 4 - 10).

In 1979 the President’s Commission on Foreign Language and International Studies presented its report
"Strength through Wisdom: A Critique of U.S. Capability" to President Carter. Among its recommendations was the setting up of a National Criteria and Assessment Program to develop language tests and assess language learning in the U.S., citing the work of the FSI as a valuable step in this direction.

ACTFL was given the role of producing National criteria, and the ACTFL Provisional Proficiency Guidelines appeared in 1982 (ACTFL, 1982) and the complete Guidelines in 1986 (ACTFL, 1986).

This gave birth to what has now come to be termed the "Proficiency Movement", and has generated equal amounts of enthusiasm among its supporters and despair among its critics. However, before considering the implications of the Proficiency Movement for oral testing and the uses to which the ACTFL rating scales have been put, it is necessary to backtrack to some degree and look at the assessment scales which were developed by the FSI and plot their development to the 1986 ACTFL rating scale, and consider the evidence upon which researchers have claimed that the scales are reliable and valid.

1.3 FSI to ACTFL: The development of oral rating scales

1.3.1 The FSI rating scale

The first FSI oral interviews were conducted by a native speaker interviewer who was selected on the criteria of "friendliness" and "interest" to elicit the speech sample, and a rater who need not be fluent in the
language, but who was sensitive to errors which the student might make (Wilds, 1979). The interview would last between 10 and 40 minutes. The procedure used was for the interviewer to begin with simple social formulae, and from there to proceed to elicit grammatical structures from the candidate. After this, a role-play or a prepared dialogue would be used. Appropriate elicitation techniques were prepared in advance. The rater would take notes, and grade the candidate on a six-point scale across five factors, as presented in Figure 2.1.

Fig. 2.1. FSI rating factors.

<table>
<thead>
<tr>
<th></th>
<th>foreign</th>
<th>:</th>
<th>:</th>
<th>:</th>
<th>:</th>
<th>:</th>
<th>native</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td>inadequate</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>adequate</td>
</tr>
<tr>
<td>Fluency</td>
<td>uneven</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>even</td>
</tr>
<tr>
<td>Comprehension</td>
<td>incomplete</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>complete</td>
</tr>
</tbody>
</table>

A weighted scoring system was devised by Wilds which was derived from the multiple correlation between each of the factors and the overall rating assigned. The weighted score was then converted into one of the FSI levels. This process (ETS, 1970) is illustrated in Figures 2.2 and 2.3.
It is clear from Figure 2.2 that grammatical accuracy is given the highest weighting in deriving the final scores; it is often claimed that this is still the case with the most recent ACTFL scales, and is the point of some contention between ACTFL proponents and their critics (Van Patten, 1986). The components which are chosen for criteria in an oral rating scale and the importance of each of those components is an area of research in which little progress has been made, apart from the work done by Adams (1980). However, in 1958 when descriptors were for the first time written to give
content to the levels of the FSI scale, a process was begun which provided the opportunity to conduct research into the relationship between the test and actual language production.

The 1958 FSI band descriptors are given here in full.

**Level 1: Elementary Proficiency.** Able to satisfy routine travel needs and minimum courtesy requirements.

Can ask and answer questions on topics very familiar to him; within the scope of his very limited language experience can understand simple questions and statements, allowing for slowed speech, repetition or paraphrase; speaking vocabulary inadequate to express anything but the most elementary needs; errors in pronunciation and grammar are frequent, but can be understood by a native speaker used to dealing with foreigners attempting to speak his language; while topics which are "very familiar" and elementary needs vary considerably from individual to individual, any person at Level 1 should be able to order a simple meal, ask for shelter or lodging, ask and give simple directions, make purchases, and tell time.

**Level 2: Limited Working Proficiency.** Able to satisfy routine social demands and limited work requirements.

Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events, as well as work, family and autobiographical information; can handle limited work requirements, needing help in handling any complications or difficulties; can get the gist of most conversations on non-technical subjects (i.e., topics which require no specialized knowledge) and has a speaking vocabulary sufficient to express himself simply with some circumlocutions; accent, though often quite faulty, is intelligible; can usually handle elementary constructions quite accurately but does not have thorough or confident control of the grammar.
Level 3: Minimum Professional Proficiency. Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social, and professional topics.

Can discuss particular interests and special fields of competence with reasonable ease; comprehension is quite complete for a normal rate of speech; vocabulary is broad enough that he rarely has to grope for a word; accent may be obviously foreign; control of grammar good; errors never interfere with understanding and rarely disturb the native speaker.

Level 4: Full Professional Proficiency. Able to use the language fluently and accurately on all levels normally pertinent to professional needs.

Can understand and participate in any conversation within the range of his experience with a high degree of fluency and precision of vocabulary; would rarely be taken for a native speaker, but can respond appropriately even in unfamiliar situations; errors of pronunciation and grammar quite rare; can handle informal interpreting from and into the language.

Level 5: Native or Bilingual Proficiency. Speaking proficiency equivalent to that of an educated native speaker.

Has complete fluency in the language such that his speech on all levels is fully accepted by educated native speakers in all its features, including breadth of vocabulary and idiom, colloquialisms, and pertinent cultural references.

It should be noted that the FSI band descriptors were developed in such a way that they could be applied to all languages; in other words it was considered to be a "common measure" (Clark, 1980a), and this explained the very real problems in ensuring that an S-3 in French was not more difficult to obtain than in other languages, as reported by Sollenberger (1978).
The scale ranged from no proficiency whatsoever to native speaker ability. The concept of the "native speaker" as an assessment criterion has affected most oral rating scales since the FSI, and is one to which we shall return.

The assessment carried out was global, but despite Sollenberger's view that there was no attempt to isolate components of speaking ability, we have seen that component scales were developed and a weighting system constructed. The component scales, given in full below, were used only in the training of FSI raters, and in live tests as a check on the accuracy of the global score awarded. However, there is no published evidence to suggest whether this practice was successful or not.

**Accent**

1. Pronunciation frequently unintelligible.
2. Frequent gross errors and a very heavy accent make understanding difficult, requires frequent repetition.
3. "Foreign accent" requires concentrated listening and mispronunciations lead to occasional misunderstanding and apparent errors in grammar or vocabulary.
4. Marked "foreign accent" and occasional mispronunciations which do not interfere with understanding.
5. No conspicuous mispronunciations, but would not be taken for a native speaker.
6. Native pronunciation, with no trace of "foreign accent".

**Grammar**

1. Grammar almost entirely inaccurate except in stock phrases.
2. Constant errors showing control of very few major patterns and frequently preventing communication.
3. Frequent errors showing some major patterns uncontrolled and causing occasional irritation and misunderstanding.

4. Occasional errors showing imperfect control of some patterns but no weakness that causes misunderstanding.

5. Few errors, with no patterns of failure.

6. No more than two errors during the interview.

**Vocabulary**

1. Vocabulary inadequate for even the simplest conversation.

2. Vocabulary limited to basic personal and survival areas (time, food, transportation, family, etc.).

3. Choice of words sometimes inaccurate, limitations of vocabulary prevent discussion of some common professional and social topics.

4. Professional vocabulary adequate to discuss special interests; general vocabulary permits discussion of any non-technical subject with some circumlocutions.

5. Professional vocabulary broad and precise; general vocabulary adequate to cope with complex practical problems and varied social situations.

6. Vocabulary apparently as accurate and extensive as that of an educated native speaker.

**Fluency**

1. Speech is so halting and fragmentary that conversation is virtually impossible.

2. Speech is very slow and uneven except for short or routine sentences.

3. Speech is frequently hesitant and jerky; sentences may be left uncompleted.

4. Speech is occasionally hesitant, with some unevenness caused by rephrasing and groping for words.

5. Speech is effortless and smooth, but perceptibly non-native in speed and evenness.

6. Speech on all professional and general topics as effortless and as smooth as a native speaker's.

**Comprehension**

1. Understands too little for the simplest type of conversation.

2. Understands only slow very simple speech on common social and touristic topics: requires constant repetition and rephrasing.
3. Understands careful, somewhat simplified speech directed to him, with considerable repetition and rephrasing.
4. Understands quite well normal educated speech directed to him, but requires occasional repetition and rephrasing.
5. Understands everything in normal educated conversation except for very colloquial or low-frequency items, or exceptionally rapid or slurred speech.
6. Understands everything in both formal and colloquial speech to be expected of an educated native speaker.

1.3.2 Critique of the FSI rating scales

These descriptors are a mixture of linguistic and non-linguistic criteria. There are undefined degrees of accuracy to be achieved for each level in each scale, mixed with references to the types of task or the situation in which the student would be expected to be able to operate outside the test situation. This mixture of criteria is another aspect of the original FSI descriptors which has been passed on to successive generations of oral rating scales, and is frequently criticised (Matthews, 1990). If prediction from a test situation to a wider number of non-test situations is the purpose of the oral test, then it is essential that the rating scale contain descriptions of those skills or abilities which researchers can demonstrate underlie successful performance across a range of situations, or to specify tasks or situations which could be demonstrated to allow generalization to other tasks or situations. To include descriptions of the task or situation in the rating scale immediately limits the
generalizability of the oral test result to similar situations, and hence undermines one of the prime purposes of testing.

Each of the scales assumes linear development in language proficiency from zero to perfect, native-speaker speech. Each band represents an increase in the accuracy of the language used, with similar modifiers being used in each rating scale. As we are primarily concerned with accuracy and fluency (see page 115), the grammar and fluency scale will be considered in a little more detail.

The grammar scale appears to hinge around the progression of modifiers from "constant", to "frequent", "occasional" and "few" errors. At band 1 grammar is "entirely" inaccurate and at band 6 we expect "no more than 2 errors" - a specific number, the choice of which does not seem to have been justified. Nowhere is it suggested what kind of errors typify each band, and there is no indication that the scale is linked to any consideration of a sequence in which students acquire certain grammatical forms. The reference to "major patterns" in band 2 may seem to suggest that the authors of the scale did have some notion of the grammatical forms that they would expect to occur in earlier and later bands, but these are not listed or described.

The situation is similar with the FSI fluency rating scale. In band 1 we are told that "conversation is virtually impossible" while in the last band it is "as effortless and smooth as a native speaker's". The
concepts which dominate the bands between the extremes are those of speed of delivery, hesitation and "unevenness", modified by "very", "frequently" and "occasionally" in bands 2 to 4. These concepts are not defined: speed of delivery may vary considerably among non-native and native speakers of a language, and unevenness appears to be linked to "rephrasing and groping for words", something which appears to be frequent in native speaker speech (Fulcher, 1987). The nature of hesitation and its causes do not seem to have been investigated, and without a theory underlying the possible reasons for hesitation it would seem unlikely that the concept could be applied in such a linear fashion to non-native learners.

As the FSI provides the model which many other rating scales have adopted, these concepts, despite being poorly defined, have found their way into most other oral rating scales produced since the FSI with little or no change. The problem has been that the descriptors were written without consideration of the way in which students actually speak or the way in which competence in a language develops; that is, they appear to be "a-priori" or "armchair" scales, produced by the scale writers from their intuitive judgement of how language competence develops and how this competence is used in a performance test.
1.3.3 The ILR rating scale

The ILR rating scales were produced in 1968 in the attempt by various Government agencies to standardize their procedures. The ILR rating scale contained 6 bands, with plus points at 0, 1, 2, 3, and 4, giving 11 possible levels of proficiency. The "Capsule Characterisations" of these levels is provided by Lowe (1985a: 14) as follows:

**Level 0:** No functional ability

**Level 0+:** Operates with memorized material, and short lists.

**Level 1:** Survival. Consistently creates with the language. Tends to operate in the present time. Uses short sentences.

**Level 2:** Concrete. Operates in past and future as well as present time. Can use paragraphs (combining sentences into limited connected discourse).

**Level 3:** Abstract. Thinks in the target language. Supports opinions and hypothesizes. Handles unfamiliar situations and topics. Can use organized discourse. General vocabulary is controlled with only sporadic errors in basic grammar, but with some errors in frequent complex structures and more errors in low-frequency complex structures. Can use colloquial and careful formal speech.

**Level 4:** Representation. Tailors language to suit the audience. Discourse is well organized and a controlled use of synonyms.

**Level 5:** Educated Native. Functions equivalent to a well-educated native speaker. Absolutely appropriate discourse in a wide range of contexts from colloquial to careful formal speech.

The scale begins with no ability and ends with native-speaker ability, although there are a number of additional assumptions made: that present tenses are used
before past and future, that control of discourse begins with the ability to use paragraphs (band 2) - although what a "paragraph" is in speech is not explained, that the ability to use colloquial and formal speech is a feature of band 3 - although we are not told which of these is used prior to band 3, and finally that awareness of audience and the puzzling "use of synonyms" is a feature of band 4. As in the FSI scale, precise definition of terms is avoided, although there is some indication as to why this is the case. Lowe (1985a: 15) claims that the ILR rating scale was developed on the basis of Gestalt psychology, arguing metaphorically that one does not have to test the soil to tell that you are on a mountain: one need only rely on the overall shape of the terrain. At least one question is whether or not the ILR descriptors do actually describe the terrain!

1.3.4 The ACTFL/ETS rating scale

The ACTFL oral rating scale, described by Lowe (1987) was specifically designed with a larger number of levels in order to discriminate more accurately between students in non-government settings at the level below 2/2+ following the studies by Carroll (1967) and the ETS (see page 15 above). Figure 2.4 on page 28 is taken from Lowe (1983: 232) and demonstrates how the ACTFL levels relate to the ILR rating scale.

The categories at the lower end of the continuum have been expanded to increase the amount of
discrimination possible, while at the higher end all the categories from S-3 upwards have been collapsed into the single category "superior".

**Fig. 2.4. A comparison of the ILR and ACTFL rating scales.**

<table>
<thead>
<tr>
<th>ILR</th>
<th>ACTFL</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Superior</td>
</tr>
<tr>
<td>4+</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>3+</td>
<td>Advanced Plus</td>
</tr>
<tr>
<td>3</td>
<td>Advanced</td>
</tr>
<tr>
<td>2+</td>
<td>Intermediate High</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1+</td>
<td>Intermediate Mid</td>
</tr>
<tr>
<td>1 (strong)</td>
<td></td>
</tr>
<tr>
<td>1 (weak)</td>
<td>Intermediate Low</td>
</tr>
<tr>
<td>0+</td>
<td>Novice High</td>
</tr>
<tr>
<td>0 (strong)</td>
<td></td>
</tr>
<tr>
<td>0 (weak)</td>
<td>Novice Low</td>
</tr>
</tbody>
</table>

The descriptors used to define the ACTFL levels are given here in full from the ACTFL Guidelines (1986).

**Novice:** The Novice level is characterized by an ability to communicate minimally with learned material.

**Novice Low:** Oral production consists of isolated words and perhaps a few high-frequency phrases. Essentially no functional communicative ability.

**Novice Mid:** Oral production continues to consist of isolated words and learned phrases within very predictable areas of need, although quantity is increased. Vocabulary is sufficient only for handling simple, elementary needs and expressing basic courtesies. Utterances rarely consist of more than two or three words and show frequent long pauses and repetition of interlocutor's words. Speaker may have some difficulty producing even the simplest utterances. Some Novice-Mid
speakers will be understood only with great difficulty.

Novice High: Able to satisfy partially the requirements of basic communicative exchanges by relying heavily on learned utterances but occasionally expanding these through simple recombinations of their elements. Can ask questions or make statements involving learned material. Shows signs of spontaneity, although this falls short of real autonomy of expression. Speech continues to consist of learned utterances rather than of personalized, situationally adapted ones. Vocabulary centers on areas such as basic objects, places, and most common kinship terms. Pronunciation may still be strongly influenced by first language. Errors are frequent and, in spite of repetition, some Novice-High speakers will have difficulty in being understood even by sympathetic interlocutors.

Intermediate: The intermediate level is characterized by an ability to
- create with the language by combining and recombin- ing learned elements, though primarily in a reactive mode;
- initiate, minimally sustain, and close in a simple way basic communicative tasks; and
- ask and answer questions.

Intermediate Low: Able to handle successfully a limited number of interactive, task-oriented social situations. Can ask and answer questions, initiate and respond to simple statements, and maintain face-to-face conversation, although in a highly restricted manner and with much linguistic inaccuracy. Within these limitations, can perform such tasks as introducing self, ordering a meal, asking directions, and making purchases. Vocabulary is adequate to express only the most elementary needs. Strong interference from native language may occur. Misunderstandings frequently arise, but with repetition, the Intermediate-Low speaker can generally be understood by sympathetic interlocutors.

Intermediate-Mid: Able to handle successfully a variety of uncomplicated, basic communicative tasks and social situations. Can talk simply about self and family members. Can ask and answer questions and participate in simple conversations on topics beyond the most immediate needs; e.g., personal history and leisure-time activities. Utterance length increases slightly, but speech may continue to be characterized by frequent long pauses, since the smooth incorporation of even basic conversational strategies is often hindered as the speaker struggles to create appropriate language forms. Pronunciation may continue to be strongly influenced by first language and fluency may still be strained. Although misunderstandings still arise, the Intermediate-Mid speaker can generally be understood by sympathetic
interlocutors.

Intermediate High: Able to handle successfully most uncomplicated communicative tasks and social situations. Can initiate, sustain, and close a general conversation with a number of strategies appropriate to a range of circumstances and topics, but errors are evident. Limited vocabulary still necessitates hesitation and may bring about slightly unexpected circumlocution. There is emerging evidence of connected discourse, particularly for simple narration and/or description. The Intermediate-High speaker can generally be understood even by interlocutors not accustomed to dealing with speaking at this level, but repetition may still be required.

Advanced: The Advanced level is characterized by an ability to:
- converse in a clearly participatory fashion;
- initiate, sustain, and bring to closure a wide variety of communicative tasks, including those that require an increased ability to convey meaning with diverse language strategies due to a complication or an unforeseen turn of events;
- satisfy the requirements of school and work situations;
- narrate and describe with paragraph-length connected discourse.

Advanced: Able to satisfy the requirements of everyday situations and routine school and work requirements. Can handle with confidence but not facility complicated tasks and social situations, such as elaborating, complaining, and apologizing. Can narrate and describe with some details, linking sentences together smoothly. Can communicate facts and talk casually about topics of current public and personal interest, using general vocabulary. Shortcomings can often be smoothed over by communicative strategies, such as pause fillers, stalling devices, and different rates of speech. Circumlocution which arises from vocabulary or syntactic limitations very often is quite successful, though some groping for words may still be evident. The Advanced level speaker can be understood without difficulty by native interlocutors.

Advanced Plus: Able to satisfy the requirements of a broad variety of everyday, school and work situations. Can discuss concrete topics relating to particular interests and special fields of competence. There is emerging evidence of ability to support opinions, explain in detail, and hypothesize. The Advanced-Plus speaker often shows a well-developed ability to compensate for an imperfect grasp of some forms with confident use of communicative strategies, such as paraphrasing and
circumlocution. Differentiated vocabulary and intonation are effectively used to communicate fine shades of meaning. The Advanced-Plus speaker often shows remarkable fluency and ease of speech, but under the demands of Superior-Level, complex tasks, language may break down or prove inadequate.

Superior: The Superior level is characterized by an ability to
- participate effectively in most formal and informal conversations on practical, social, professional, and abstract topics; and
- support opinions and hypothesize using native-like discourse strategies.

Superior: Able to speak the language with sufficient accuracy to participate effectively in most formal and informal conversations on practical, social, professional, and abstract topics. Can discuss special fields of competence and interest with ease. Can support opinions and hypothesize, but may not be able to tailor language to audience or discuss in depth highly abstract or unfamiliar topics. Usually the Superior level speaker is only partially familiar with regional or other dialectical variants. The Superior level speaker commands a wide variety of interactive strategies and shows good awareness of discourse strategies. The latter involves the ability to distinguish main ideas from supporting information through syntactic, lexical, and suprasegmental features (pitch, stress, intonation). Sporadic errors may occur, particularly in low-frequency structures and some complex high-frequency structures more common to formal writing, but no patterns of error are evident. Errors do not disturb the native speaker or interfere with communication.

The ACTFL scale is marked by an increase in the amount of information which is provided for the rater, but the lack of detailed explanation of the terms used or potential exponents in actual speech continues to be a mark of the prose descriptions. No empirical evidence is available to confirm that new criteria introduced into the rating scale such as "discourse", "interactive" or "communicative" strategies do discriminate between the students at the proposed ability levels where these
occur. Indeed, Lowe (1985a: 16) states that "the use of the system remains implicit" which seems to imply that the descriptors as they stand must be interpreted by each reader, although rater training is essential for any practical use of the scale. Finally, we make the observation that the mixture of linguistic and non-linguistic criteria is increased in the ACTFL scale, indicating task type and topic area which may be dealt with at a given level of ability. This confusion of criteria has been severely criticised (Bachman and Savignon, 1986; Matthews, 1990) as it makes validation studies extremely difficult. We will return to this issue when looking at validation procedures below.

1.4 Discussion

Having described the development of, and presented an introductory critique of the FSI, ILR and ACTFL rating scales, the basis for most other oral rating scales, the main failings may be identified. Perhaps most important is the fact that none of the scales have any "empirical underpinning" (Lantolf and Frawley, 1985; Pienemann et. al., 1988: 218). Indeed, Lowe (1987: 46) admits that there have been few studies into the reliability and validity of the Oral Proficiency Interview and the rating scales upon which the scores are awarded.

Investigation into the way in which students acquire language, in terms of competence or performance, has not been taken into account in the construction of these

They argue that:

"Such descriptions are so vague and general as to be utterly unhelpful in distinguishing any second language learner from another. If "areas of weakness" can be construed to mean areas in which learners' usage does not conform to the standard, then every language learner conforms to this description. Numerous research studies have shown that learners do not suddenly "learn" a structure and begin to use it correctly 100% of the time....Even the most advanced of second language learners will therefore display "weaknesses" in the areas cited."

Similarly, Valdman (1988: 121) argues that:

"...it is fair to say that although the OPI (Oral Proficiency Interview) may be experientially based, its theoretical underpinnings are shaky and its empirical support scanty."

It is thus argued that without a sound empirical basis for initial rating scale development it makes little sense to investigate the validity of an oral rating scale post-hoc when results cannot be related to initial hypotheses and constructs, and other confounding factors such as rater training have been introduced. We agree with Jarvis (1986: 21) when he argues that:
"After-the-fact inquiry is unacceptable and has historically degenerated into little more than validation of flawed systems."

In terms of the actual language produced by students which could be classified at certain levels on the rating scale, one example from the literature will be considered briefly. This particular example is not a transcript of an actual interview, but a constructed dialogue to exemplify the difference between a Novice level conversation and an Intermediate level conversation taken from Liskin-Gasparro (1984a: 480).

Novice Level Conversation:
Teacher: Tell me about your family.
Student: (Silence)
Teacher: How many people are in your family?
Student: Four.
Teacher: Who are they?
Student: Mother. Father. Brother. Me.
Teacher: Tell me something about your brother.
Student: (Silence)
Teacher: How old is your brother?
Student: 14.
Teacher: What is his name?
Student: John.

Intermediate Level Conversation:
Teacher: Tell me about your family.
Student: I have four people in my family. I have a mother. I have a father. I have a brother.
Teacher: Tell me about your brother.
Student: My brother's name is John. He is 14. He plays football. He plays the violin.

It is claimed that this second example shows that the student has "crossed the all important threshold from
operating with memorised material and isolated words and phrases at the Novice level to creating with language at the Intermediate level" (Liskin-Gasparro, 1984a: 481).

This is not particularly enlightening, as constructed conversations produced post-hoc to justify a-priori rating scales in no way provide evidence to suggest that the rating scale descriptors provide valid criteria to discriminate between students at differing levels of ability. It could be countered that this is merely an example and is relatively unimportant. However, as it is the only example available from the published literature on the ACTFL rating scale, and no examples of student speech appear to have been analyzed before the rating scales were constructed and operationalized, it would seem that this criticism is justified.

A further justification of the ACTFL rating scale descriptors is found in the exposition of the Functional Trisection which was designed to provide the scale with a functional/notional emphasis (Liskin-Gasparro, 1984b: 35). The Functional Trisection has also been used by other researchers in oral testing, notably Adams et. al (1987) and Griffin et. al. (1988). However, the Functional Trisection provides further evidence for the lack of an empirical basis to oral rating scale development in the confusion it generates as a post-hoc rationale for the ACTFL rating scale. It is claimed that each level in the ACTFL rating scale contains statements about what the students can do in three areas: (1) the
linguistic functions or tasks that a candidate can perform, (2) the context, or topics that can be handled, and (3) the degree of accuracy with which the message will be communicated or understood. In order to give content to these three "functions" five linguistic factors are taken into account: grammar, vocabulary, fluency, pronunciation and sociocultural ability. Whether or not the last of these is a "linguistic" factor is a matter for debate, but this is what is claimed. The Functional Trisection for levels within the ACTFL rating scale is given below.

**Level 0** (Novice categories are not separated)

Function: No functional ability.
Context: None.
Accuracy: Unintelligible.

**Level 1** (Intermediate)

Function: Can create with the language, ask and answer questions, participate in short conversations.
Context: Everyday survival topics and courtesy requirements.
Accuracy: Intelligible to native speaker used to dealing with foreigners.

**Level 2** (Advanced)

Function: Able to fully participate in casual conversations, can express facts, give instructions, describe, report, and provide narration about current, past, and future activities.
Context: Concrete topics such as own background, family interests, work, travel, and current events.
Accuracy: Understandable to native speaker not used to dealing with foreigners, sometimes miscommunicates.

**Level 3** (Superior)

Function: Can converse in formal and informal situations, resolve problem situations, deal with unfamiliar topics, provide explanations, describe in detail, offer supported
opinions and hypothesize.
Context: Practical, social, professional, and abstract
topics, particular interests, and special fields of
competence.
Accuracy: Errors never interfere with understanding and
rarely disturb the native speaker. Only sporadic errors
in basic structures.

Level 4 (Superior)
Function: Able to tailor language to fit audience,
counsel, persuade, negotiate, represent a point of view,
and interpret for dignitaries.
Context: All topics normally pertinent to professional
needs.
Accuracy: Nearly equivalent to an educated native speaker
(ENS). Speech is extensive, precise, appropriate to every
occasion with only occasional errors.

Level 5 (Superior)
Function: Functions equivalent to an educated native
speaker.
Context: All subjects.
Accuracy: Performance equivalent to an educated native
speaker.

In this description, functions are defined as the
tasks accomplished, attitudes expressed and the tone
conveyed. Context refers to topics, subject areas,
activities and jobs addressed. Accuracy is the
acceptability, quality and accuracy of the message
conveyed.

It seems from the descriptions above that there is a
serious mismatch between the Functional Trisection and
the oral rating scale it was designed to defend. The
Novice descriptions (especially Novice High) do not
correspond with either the description in the rating
scale or the example of constructed speech provided by
Liskin-Gasparro to exemplify the kind of language which
might be expected from a student at this level. Further, the Superior level is broken down into three sections within the Functional Trisection implying that these represent a linear development in language proficiency, whilst in the rating scale the Superior category is not broken down into subcategories. It would therefore appear that the Functional Trisection as a post-hoc theoretical justification of the ACTFL rating scale is largely irrelevant.

Despite the problems which have been discussed above, the model of language learning assumed by the ACTFL/ETS/ILR rating scales (often shortened to AEI), in which very little progress in learning the language is needed to gain a higher band in the early levels and much more progress needed to register improvement in the later levels (Lowe, 1985a: 21), has become the basis for a whole approach to language teaching and testing. This is known as the Proficiency Movement, and the wide acceptance of the principles of the movement constitute a strong claim for the validity of the AEI rating scales.

Thus, Lowe (1987: 47) claims that

"Its [the AEI model of language acquisition as represented in the band descriptors of the rating scale] ultimate utility may lie beyond testing per se in its effect on curriculum. In this case, teaching for the test - teaching for general functional foreign language ability - is not to be discouraged."

38
The strong form of this argument is that "testing drives curriculum" (Lowe, 1983: 238 - 239). Similarly, Liskin-Gasparro (1984b: 35) claims that the ACTFL Guidelines are the first step in "creating proficiency-based curricula that respond to the need of the profession to define what students should be able to do with the language at various stages in the learning process."

However, limits to the usefulness of teaching to the test and positive washback effect on curriculum have been set. While accepting the importance of the ACTFL Guidelines for oral testing, Brown (1985) has argued that the ACTFL global rating scale cannot be used for evaluation in the classroom as it does not provide adequate informative feedback to the learner or the student. Similarly, Liskin-Gasparro (1984a: 486) writes that:

"...it is important to remember that the ACTFL Guidelines are not curriculum outlines, nor are they prescriptions for what grammatical structures to teach and when. They are a graduated sequence of proficiency states around which a foreign language program may be structured. The day-to-day activities that constitute the sequence of small steps in the context of the larger phases identified by the Guidelines, as well as the methods to be used, are still and always will be the province of the foreign language teacher."
Although limits are placed on the usefulness of the Guidelines in the classroom, this latter quotation amounts to a claim that the AEI band descriptors in the oral rating scales represent the way in which students do actually acquire language, and are valid from this perspective (Barnwell, 1987; Pienemann et. al. 1988). The wide acceptance of this claim has meant that language courses are often based around the ACTFL level descriptors (Magnan, 1986a: 429) and the oral proficiency interview elicitation techniques and tasks for each ability level used as teaching techniques (Magnan, 1984b). Clark (1988a) argues that washback from the AEI approach to oral testing has changed the way in which language is taught, introducing realistic communicative activities into the classroom. Once again, it must be emphasised that there is no evidence upon which to judge the claims made for the AEI rating scales, and wide acceptance of the rating scales in practice does not in any way constitute an argument for validity.

In conclusion to this section, we may claim that although the tradition of oral testing in the United States has been longer than that of other countries and has generated many of the important questions which remain to be answered by researchers, the speed with which tests have been operationalized and validated post-hoc has been somewhat unhelpful in the process of learning more about the validity of rating scales and their relationship to any hypothesized reality external
to themselves.

In Section 2 we turn to look at what research has been conducted into reliability and validity of oral rating scales, assessing both the tradition of the United States and work carried on in other countries where it is relevant.

Section 2. Issues in reliability and validity in oral testing.

2.1 Reliability in oral testing

2.1.1 Evidence from the United States

Lowe (1987: 46) acknowledges that there have been few studies into the reliability and validity of the Oral Proficiency Interview (OPI) and the rating scales upon which the scores are awarded, but refers to the experience of assessors with the use of the OPI and the successful placement of Government personnel over a long period of time as justification for procedures used. Lowe (1986: 394) states that

"The ILR approach has permitted successful use of the WENS (well educated native speaker) concept as the ultimate criterion in government for over thirty years."

Wilds (1975) had already put forward the same argument. Above all, the tradition of AEI evaluation must continue unbroken (Lowe, 1986: 396). The point being made is that the top band in the scale is a fixed point capable of description, to which lower bands may be
sequentially linked, and raters are capable of reliably using such a rating scale because of their understanding of this fixed criterion.

This is the experiential claim for reliability which is common to much work on oral testing both within the AEI approach and elsewhere.

One study which is constantly referred to within the literature is that of Adams (1978) on inter-rater reliability. This single study is cited as justification more often than not for the reliability of all forms of the OPI and associated rating scales, as it was found that agreement between two raters was consistently in the order of .87 or higher. Liskin-Gasparro (1984a: 483) appeals directly to Adams' study in defence of the reliability of the ACTFL, even though the raters were not using the same scales. This is surely unjustified.

However, it is generally claimed in the literature that high inter-rater reliability is not difficult to achieve in oral tests. One researcher who disagrees with this view, but provides no evidence in support of his opinion is Harrison (1982: 2). Even Barnwell (1987: 36) who is a critic of the AEI agrees that high rater reliability has been demonstrated. Barnwell (1986), for example, reports a study in which 7 teachers of Spanish were provided with "brief" training in the use of the ACTFL rating scale and then asked to rate students from video recordings of OPIs. 41% of grades were exactly the same as actual ACTFL scores, 45% were within 1 band and
14% 2 or more bands away. Barnwell concluded that this evidence shows that even untrained raters can reach acceptable levels of agreement.

Mullen (1978b) conducted a study in which two judges and four rating scales were used on two groups of students. Each judge rated both groups of students on each of the scales, and it was found that there was no significant difference in the mean scores of the two groups, the judges, or the operation of the scales. Variance was primarily attributable to the difference between students, although a small interaction effect was found between judge and scale, and judge and student. Mullen concluded that inter-rater reliability was high when two judges were used, something which is common procedure in the FSI and AEI oral tests. These findings are supported by Clark (1975: 16 - 17) and Clark and Swinton (1980a). In the latter study however, the results appear to be inconclusive. The scores of two raters on 94 students across the five factors of the FSI were used to assess inter-rater reliability. The correlation coefficients are contained in Table 2.1 in the Appendix to Chapter 2. Clark concluded that the FSI was a reliable measure of spoken English. However, it would appear that some of these correlations are unacceptably low, the only correlation to reach .8 being that on grammar. Clark’s conclusion does not seem to be justified.

Hendricks et. al. (1980) report an inter-rater reliability study in which an FSI-type procedure was
used, although no linguist was available to do the rating. 182 students from the University of Southern Illinois were tested, and taped interviews rated by two teams of raters made up from a pool of one instructor and three graduate assistants. The correlation coefficients are presented in Table 2.2 in the Appendix to Chapter 2. Here, inter-rater reliability is acceptable in all but the assessment of accent. The global score appears to be highly reliable. But it must also be noted that in this study the assessment was carried out by teams of raters working together rather than individuals, and this would probably have had the effect of inflating the correlations. As such, the results must be treated with a certain amount of caution.

From studies of inter-rater reliability Mullen (1980) argued that two raters are required in any oral test, as individual raters tend to have different patterns of rating. The low correlations in Clark and Swinton's study (1980a) are an indication of this, whereas with multiple raters Hendricks et. al. (1980) produced much higher correlations. Indeed, most investigations of reliability in oral testing recommend the use of at least 2 raters (de Charruf, 1984) in order to avoid the possible effect of the rater as test bias, and also increase reliability.

We also note that the Pearson Product Moment correlation is reported as an inter-rater reliability coefficient in all these studies, and this statistic does
not take into account rater severity. Bejar (1985) argued that although there is often substantial agreement in the sequencing of students by raters, and hence high correlation coefficients, the severity of raters differs widely. Multiple raters would therefore seem to be essential to provide a correction facility for varying rater severity.

Lowe and Liskin-Gasparro (1986a: 4) claim inter-rater reliability for the OPI to be within one plus point on the rating scale when using trained raters. In the light of the other studies reviewed here, this does not seem to be an unreasonable claim as long as multiple raters are used. A series of studies conducted by Shohamy also serve to bear this out. Shohamy (1983a) reports a study in which a team of judges were asked to rate recorded tapes of interviews with 106 students of Hebrew at the University of Minnesota. Four weeks later 32 tapes were randomly selected for re-rating. Inter-rater reliability was found to be .99. A similar study (Shohamy, 1983b: 535) found inter-rater reliability to be .93.

Barnwell (1989) conducted an inter-rater reliability study using the ACTFL rating scale to see if native speakers of Spanish reacted to foreign language proficiency in their language in the way predicted by the rating scale. Although this procedure had been suggested by Byrnes (1987) as a potential method of validating the ACTFL rating scale, Barnwell points out that the scale
was developed intuitively, and not in relation to actual native speakers, and so Barnwell hypothesized that inter-rater reliability between native speakers and ACTFL raters would be low. Barnwell taped four students doing an OPI and asked 14 native speakers of Spanish with a minimum of training in the use of the ACTFL rating scale to rate them. Barnwell found that the four students were placed in the same order of ability by all raters, but that three out of the four students were given vastly different band ratings by the native speakers and the trained ACTFL raters. The training of the ACTFL raters would appear to have influenced the way in which they rate, as predicted by Alderson (1983).

Having realised that the correlation coefficient in itself is not sufficient to claim reliability, Barnwell concludes that:

"...it appears from the present study that the operational descriptions for each ACTFL/ETS level are not in themselves self-sufficient or self-explanatory; they can mean different things to different people. The scale's limpid logic is not immediately visible to the untrained eye."

Clark (1988a: 17) refers specifically to the "fine-grained" ACTFL descriptors and wonders whether or not they have any "real world validity" in that they represent "true differences in communicative performance as far as native-speaker interlocutors are concerned". He recommends the kind of research which Barnwell has
conducted, stating that only if naive native listeners could perceive the differences in levels clearly could one have a great deal of confidence in the ACTFL band descriptors. It is noted here that Barnwell does seem to have provided some limited evidence to indicate that it may not be possible to meet the conditions which Clark has set out.

2.1.2 Evidence from other countries

A question which remains to be asked is whether or not research with tests in other countries tends to confirm the general findings of high inter-rater correlation coefficients of researchers within the AEI tradition.

Reliability studies carried out with other oral tests tend to bear out the claims of high reliability made by researchers working within the AEI approach, at least as far as the correlation coefficients are concerned. (It may of course be the case that low correlation coefficients are not published!) When reporting on the development of a new oral test for use in schools in Israel, Shohamy, Reves and Bejarano (1986: 212) state that low reliability remains a problem in oral testing, and go on to describe four experimental tests (an oral interview, a role-play exercise, a reporting task and a group discussion), providing inter-rater reliabilities as .91, .76, .81 and .73 respectively. The very high correlation coefficient for the oral interview
is not uncommon, although the reliability coefficients for the role-play exercise and the group discussion are low. A similarly high figure is reported for the ASLPR interview procedure, for example (Ingram, 1982: 26; 1985: 249).

Morrison and Lee (1985) report on an oral test which was a simulation of an academic tutorial developed at Hong Kong University. The tests were videotaped and each student rated by a mixture of native and non-native speakers on a 5 band scale for the factors of general proficiency, use of grammar, pronunciation, fluency, communicative ability and academic potential. Inter-rater reliability was quoted as .85 with no significant difference between the rating behaviour of native and non-native raters.

Jafarpur (1988) used an FSI-type oral interview at Shiraz University, Iran, on which 3 judges rated 58 students on the five FSI factors. Inter-rater reliability was reported as lying between .58 and .65, and Jafarpur concluded that there was a significant difference between the rating patterns of the 3 judges. Although this result differs from that of Morrison and Lee, it must be said that the correlation coefficient provided by the latter was for EFL teachers whilst the results reported by Jafarpur are for subject teachers. This fact could explain the different results.

However, we may conclude from the published evidence regarding inter-rater reliability that this form of
reliability is fairly easy to achieve when using multiple trained raters, whereas reliability figures appear to decrease in magnitude when untrained raters and untrained naive native raters are used.

2.1.3 Reliability and rater training

Barnwell (1989) suggests that the phenomenon of higher inter-rater reliability coefficients when using trained raters and lower coefficients when using untrained raters can be accounted for by the fact that trained raters undergo a period of "socialization" during which they begin to see language proficiency in terms of the scale which they are using. By constant operation of the scale a high degree of reliability is achieved. Wilkinson (1968: 126) had previously argued in a similar rather pessimistic vein that high inter-rater reliability could only be achieved when the assessors hold similar views about the nature of speaking. The process of rater training, the socialisation into a common view of spoken language, is described by Alderson (1991b: 64) as "cloning."

It would appear from this discussion that Bachman's view (1988: 150) that the issue of reliability in oral tests is not a problem is unjustified. It is a problem if you do not have it, and it is a problem for validity studies if you have it at the expense of rater training, or "cloning". That is, rater training assumes the scale to be valid and that training raters to use it in live
testing situations is legitimate. Validity must, however, be established before raters are trained to use a rating scale. Failure to establish validity prior to rater training would introduce the training itself as a test method facet which could not be separated from the trait which the rating scale is supposed to measure. The fusion of trait and test method facet in this case would preclude the possibility of convincing post-hoc validation, whether the rating scale were valid or not.

2.1.4 A methodological note on reliability studies: some recent developments

The studies which have been reviewed above have relied, as has already been noted, on correlation coefficients. These do not take into account rater harshness, as noticed in the Barnwell study (1989), and therefore high correlation coefficients cannot be used as sufficient evidence to claim reliability (Krzanowski and Woods, 1984). A number of recent studies have used much more powerful methods to investigate inter- and intra-rater reliability, particularly Rasch Partial Credit models. Multi-faceted models in particular (Linacre, 1991) provide tools which can fully take advantage of developments in the theory of test method facets (Bachman, 1990). Many of the limitations of previous reliability studies can now be overcome.

An example of recent research which has attempted to overcome these methodological limitations is that of Lunz
and Stahl (1990a), who examined the scores given to 270 medical students by 46 judges on 3 tests, one of which was an oral interview. The oral test was conducted on three separate occasions. The computer program Facets (Linacre, 1990) was used to analyze the data, which revealed variation in judge severity and an interaction effect between judge and time of interview. However, fit statistics did show that judges were generally internally consistent despite this variation. Lunz and Stahl (1990b) also used the same method to investigate the ratings of 293 students by 48 judges on an oral test, and found that whilst there was wide variety between judges, each judge was internally consistent. This finding tends to confirm the results of Bejar (1985), and provides supporting evidence for the view that if intra-rater reliability can be established, lack of inter-rater reliability can be dealt with in other ways (Alderson: 1991b).

2.1.5 Conclusions

From the work reviewed on reliability in oral testing a number of important issues have emerged.

Firstly, work on reliability suffers from methodological limitations such as the reliance on correlations which were often not recognized when the studies were conducted. It is therefore important that in any future consideration of reliability new tools and methods must be used, especially tools which may account for the influence of test method facets on any
consistencies noted in the correlation coefficients.

Secondly, high inter-rater reliability appears to depend upon the degree of rater training or raters sharing views of the nature of language proficiency. The price to be paid for achieving high reliability through rater training is to make validation studies of rating scales impossible, as validity is assumed by rater training. It is therefore important that raters are not trained when rating scales and testing procedures are being developed and their validity examined, but only after validity studies have been carried out. Rater training would be introduced only immediately prior to operationalization.

Thirdly, the temptation to assume reliability and begin using a testing system prior to the appropriate reliability and validity studies being carried out should be avoided.

2.2 Validity in oral testing

It has been claimed that there is a tension between reliability and validity (Underhill, 1982). Achieving high reliability in a test is said to be dependent upon using a rating procedure which is as close to the objectivity of the "correct answer" in a well designed multiple choice question as possible. In this case the rating procedure is mechanistic and does not involve the personal judgement of the rater at all. On the other hand, more open tasks such as an oral interview depend
upon rating scales which involve the judgement of a rater, and scores may vary between raters (inter-rater reliability) and between the scores awarded by a particular rater to a given student at different times (intra-rater reliability). The argument appears to be that the higher the reliability through control of possible responses and the rating procedure, the less valid the test, whereas the higher the validity through creating a "realistic" or "authentic" task the lower the reliability of the rating scale and its use.

In this study, the view is taken that this is a false polarity. It has become axiomatic in the testing literature that reliability is a prerequisite to validity (Bachman and Palmer, 1982: 458; Stevenson, 1981: 47), although Davies (1990: 22 - 23) warns that consistency in itself is "trivial" if the consistency is a function of test method facets. What matters is that the reliability coefficient is an accurate estimation of the consistency with which a trait is being measured. Thus, if the oral rating scales developed in this study turned out to have low reliability coefficients, it would make little sense to proceed with validation studies as the rating scales could not be used to measure anything. If, however, the rating scales are seen to be reliable, the validity studies may be undertaken to assess the degree to which those rating scales can be seen to be measuring the intended trait. Rather than reliability and validity operating in some kind of polar relationship to each
other, it is argued here that they are both essential complementary ways of asking important questions about oral rating scales.

2.2.1 **Validity claims and some empirical evidence.**

The issue of validity in relation to the FSI and AEI oral rating scales illustrates most of the problems which are currently faced by researchers in the field of oral testing. As in the case of reliability, most of the available published work on validating oral rating scales has been done within the FSI and AEI tradition. However, it will be seen in what follows that much of the discussion centres around the face validity of the actual testing procedure. In this, it must be remembered that there is a clear implication that the rating scales (and the band descriptors) used in the testing procedure are valid in that they accurately describe the speaking ability of the students.

The main argument presented in support of the validity of the FSI testing procedure and its associated rating scale was that it was valid on the grounds that no other testing instrument existed which met the particular needs for which it was developed (Wilds, 1979: 11 - 12). Wilds maintained that the validity of the FSI was "unquestionable" because the oral interview was based upon a demonstration of speaking ability in a "natural context" related to living and working abroad, and that the rating scale used in the scoring procedure described
that speaking ability. Face validity, defined as "the degree to which students feel they are performing a real communicative act" (Bartz, 1979), was thereby pushed to the fore in discussions of validity.

The claim to validity, which is essentially face validity, has three related aspects: (a) the oral (interview) test is direct (Clark, 1975: 14); (b) it involves a natural context and real life tasks which produce natural conversation; (c) the rating scale describes distinguishable levels of speaking ability demonstrated in the test. The first of these three aspects has been severely criticised on the grounds that no test can be direct: any test aims to elicit a speech sample which provides evidence of underlying competence or ability (Bachman and Savignon, 1986: 382 - 383). The second aspect has been questioned on the basis of studies of "interview talk" which suggest that the interview generates a special "genre" of language different from normal conversational speech (Silverman, 1976; MacPhail, 1985; van Lier, 1989; Perrett, 1990). The third assumption, that the rating scale does describe distinguishable levels of speaking ability, has been questioned on the basis of analysis of actual speech in comparison to band descriptors (Fulcher, 1987; Fulcher 1989; Perrett, 1990; Matthews, 1990) and the scalability of the bands of the rating scale (Pienemann et. al. 1988).

However, the basic position with regard to claims of
validity in oral tests does not seem to have changed greatly, and proponents of so-called "communicative" language tests still make the notion of "real-life tasks" and "natural language" the touchstone of validity, with the assumption that their rating scales do adequately describe natural language use (Morrow, 1982: 56 - 57; Underhill, 1987). In fact, many writers merely argue that the use of real-life tasks ensures the use of natural language, and assume that the rating scale in use in the scoring process does adequately describe natural language use. A claim to face validity is often all that is used to defend an entire testing system including its rating scale.

For example, in his review of the ILR Lowe (1987: 46) claims to discuss issues of reliability, validity and practicality. The single paragraph devoted to validity merely claims high face validity because it requires examinees to use the spoken language. Barnwell (1987: 36) goes as far as to say that the FSI and its descendants rely entirely on claims of face validity. It is not surprising to find that testers such as Stevenson (1981; 1985a; 1985b) who reject the use of face validity as the sole evidence for evaluating oral tests are criticised by those who espouse a "common-sense" non-technical approach to oral testing (Underhill, 1987).

Lowe (1983: 235) argues that the OPI maintains content validity because the questions posed by the interviewer reflect real-life situations, whereas Lowe
and Liskin-Gasparro (1986: 4) claim that the OPI is highly face valid because it involves real conversation and the rating scales have content validity because of the use of descriptors for each proficiency level within the rating scale. This statement is not fully explained, but we may take it that it represents a claim to the psychological reality of the scale. As we have seen, Pienemann et. al. (1988) and Barnwell (1987; 1989) have cast doubt upon this claim.

However, the most common defence for the validity of an oral test and its rating scale is that of experience. Liskin-Gasparro (1984a: 483) uses this argument, but its strongest expression is found in Lowe (1986: 392) written in response to criticisms by Kramsch (1986b) and Bachman and Savignon (1986). Lowe writes:

"The essence of the AEI proficiency lies not in verbal descriptions of it, but in its thirty-year-long tradition of practice - making training in AEI proficiency testing a desideratum."

Lowe proceeds to argue that those who question the validity of the AEI rating scales do so because they are not among the AEI "adepts". It is only they who have been trained in the system and understand the words on the paper. Those who have not been trained are incapable of understanding the rating scale.

In the preceding discussion it can be seen that two themes run through the debate. One is the validity of the
interview itself, that is, the tasks which the students are asked to do. The second is the claim that the rating scale, the scoring part of the process, is a valid description of levels of ability. Similar arguments are produced for both aspects of the testing process. We wish to look at what empirical evidence has actually been presented regarding the second issue concerning the rating scales, which is the primary focus of this study.

Magnan (1986a) and Meredith (1990) go someway toward providing evidence for validity by group discrimination for the AEI oral testing procedure which, whilst important (Clark, 1988a: 17 - 18), is a weak form of validity which does not provide sufficient evidence in itself for any strong claims (see pages 315 - 316).

Magnan (1986a) describes a study carried out at the University of Wisconsin to investigate the relationship between the level of study and the level of proficiency of French students as measured by the ACTFL rating scale, which was essentially an investigation of validity by group differences. 40 students, 10 from each level of the French program, were tested by a certified ACTFL interviewer and the interviews taped and rated by two independent raters from the ETS. Inter-rater reliability was reported as .72, which was not as high as Magnan had hypothesized it would be.

Magnan discovered some overlap between ACTFL band scores and level of study, which is presented in Table 2.3 in the Appendix to Chapter 2. However, this overlap
could be accounted for by the amount of unreliability in the testing process. Magnan (1986a: 431 - 432) justifiably concludes that there is, on the whole, a general relationship between years of study and proficiency ratings, except for the third and fourth year, where she argues that the ACTFL scale is not sensitive enough at the higher levels.

Meredith (1990) repeated Magnan's study using students of Spanish at Brigham Young University, and concluded that the ACTFL rating scale, as it stands, is "a very blunt instrument." After trying to make the rating scale more sensitive to years of language study, Meredith could only report a highest correlation of .57 between ACTFL ratings and length of study (ibid., 294).

It could of course be the case that in these studies the assumption that students within each year are equal in ability is unjustifiable, and would have led to the negative findings. Similarly, Meredith assumes that proficiency increases in a linear fashion with length of study, which is not obviously the case.

In a similar study, Adams (1980) attempted to investigate the relationship between the five factors of the FSI rating scale (fluency, comprehension, grammar, vocabulary and accent) and the final rating of the students. The ratings of 834 tests conducted during the summer of 1978 in 33 languages were used in the study, which employed discriminant analysis in order to investigate the extent to which any factor was
significantly contributing to the final score at any
given proficiency level.

Adams' results are presented in Table 2.4 in the
Appendix to Chapter 2, which shows which of the five
factors discriminate effectively at each level of the FSI
rating scale. Although from Adams' results it would seem
that different factors do appear to be playing different
roles at each level of the scale, it is almost impossible
to relate the results theoretically to the FSI level
descriptors, as the descriptors are simply too vague to
allow this sort of analysis, especially as there does not
appear to be any pattern to the factors which
discriminate at different levels. However, the idea that
discriminant analysis could be used in this way once the
description of the data is specific enough may be a
useful way of investigating the validity of rating
scales, or in rating scale construction. Multivariate
techniques of this kind do allow powerful analysis of
test data (Ross and Berwick, 1990).

Adams concludes that, as the FSI raters first record
a global score and then record the factor scores to
explain it, the study shows that at certain levels they
are more sensitive to different aspects of performance.
This is used by Adams as an argument against the use of
untrained raters as they cannot be aware of the
complexity of the rating process in determining the
proficiency of a student. Similar conclusions were
reached in a study by Clifford (1980) along the same
lines. We must conclude that although Adams' work is potentially of use to the study of validity of scoring procedures and rating scales in terms of the ideas which she generates, her particular study does not add anything to the evidence regarding the validity of the FSI because of the seemingly random distribution of the discriminatory capacity of the five factors and the impossibility of relating them theoretically to the level descriptors.

Hendricks et. al. (1980) present evidence to suggest that the various component scales of the FSI do represent independent aspects of oral proficiency, but argue that oral proficiency is a "unitary competence" which cannot be subdivided. This would lead to the conclusion that the component scales were not valid in themselves. This argument was based upon a Principal Components analysis of the ratings of two teams of judges using 64 students. Loadings on a g-factor were in the range of .79 to .87 for all factors except for accent, which was .59 (Hendricks et. al. 1980: 83).

However, it would be a mistake to take these results as having any relevance to oral rating scales in general, as the component scales are designed in such a way that if a score of, say, 2 is given on one of the scales it would be natural to give 2 on the others. It would actually be surprising if any other result had been gained, as the band descriptors encourage such cross contamination. However, there is a further reason why
this work should not be taken as presenting evidence for oral ability being a unitary factor. The research was conducted in 1976 - 1977, and was known as the Carbondale Project. This was a follow-up study to a study conducted in 1976 by Oller and Hinofotis (Oller and Hinofotis, 1980), which suggested that a g-factor could account for all variance in the speaking tests which they used. Oller (1979: 430 - 431) claimed that the research of Callaway (1980), Mullen (1980) and the Carbondale Project supported the unitary competence hypothesis (Oller, 1979: 441 - 455). The design faults of these studies have been well documented (Vollmer, 1983; Vollmer and Sang, 1983; Woods, 1983), and so the conclusions which were drawn cannot be taken as generalizable.

Of continuing relevance is the work of Bachman and Palmer (1983). They observed that much of the validation research in oral testing was concerned with correlating scores on oral rating scales with external criterion tests. The criterion was often the FSI as it was assumed to be valid. Bachman and Palmer pointed out that an assumption is not a demonstration of validity, and called for construct validation research into the FSI.

Using a Multitrait-Multimethod approach following Campbell and Fiske (1959) and confirmatory factor analysis after Joreskog (1969), Bachman and Palmer analyzed the test scores on two traits (speaking and reading comprehension) measured by three methods (the FSI oral interview, translation and self-rating).
Reliabilities for the tests ranged from .85 to .99, and convergent validity from .46 to .97. For discriminant validity, 65% of the results satisfied the criterion for speaking and 63% for reading. Method effect was found to be prominent in the translation and self-rating methods, but not in the oral interview. The most important results for the FSI were the findings that in the factor analysis it loaded most heavily on the speaking trait (.82) and had the lowest loading on the method factor (.46). Bachman and Palmer (1983: 168) concluded that

"...the oral interview maximises the effect of trait while minimising the effect of test method."

They felt confident that they had provided important evidence for the convergent and discriminant validity of the FSI. It is not without significance that this study is today quoted as one of the few which clearly provides evidence for the validity of the FSI oral interview, but is also taken, by implication, as providing evidence for the validity of the AEI rating scales and testing procedures also (Lowe, 1987: 46).

2.2.2 Concurrent validation

Concurrent validation of oral tests and oral rating scales allows investigation into (a) the relationship of speaking measures to those in other skill areas, and (b) the relationship between speaking measures.
The use of an external criterion in this way is, according to Messick (1989a: 45), a special form of construct validity. That is, it would be expected that measures which are claimed to tap the same trait will tend to converge whilst measures which are not claimed to tap the same trait will diverge. "That is, the constructs represented in the test should rationally account for the external pattern of test correlations." Convergent validity would appear to be an issue in studies where the elicitation techniques are different across two tests and the rating scale held constant, even if the aim of a piece of research is to justify the use of an indirect testing method in place of a direct method. Similarly, convergent validity of the rating scales is an issue where different rating scales are used, but where both are designed to represent the same construct. In this second case, it must be assumed that one of the rating scales has already been validated. In concurrent studies where scores on an oral rating scale are being compared with scores from a test on a different trait using the same elicitation technique, divergent validity is being investigated. However, concurrent studies which use different scoring procedures and different elicitation techniques are prone to be uninterpretable because of the confusion between traits and method facets.

Such concurrent studies have been carried out using correlational analysis, Factor analysis (Clark, 1975: 11; Anastasi, 1990, 166), and Multitrait-Multimethod studies.

However, concurrent studies of oral tests do not seem to have provided the researcher with any significant data upon which validity can be claimed. This would certainly appear to be the case with the Australian Second Language Proficiency Rating scales (Ingram, 1985b; 1990), where the ASLPR was correlated with the Comprehensive English Language Test (CELT). The CELT is a three part test consisting of listening, structure and vocabulary (Oxford, 1987) and results on this were correlated with students' scores on the ASLPR scales (Ingram, 1984).

It could very well be claimed that the CELT is an inappropriate criterion as it does not allow the collection of either convergent or discriminant evidence, but Ingram discovered a correlation between the ASLPR speaking and CELT structure of .83 when using students following courses in Australia. When using a different population of students in China the correlations were generally much lower (Ingram, 1985b: 249 - 250). It is necessary for him to interpret this in terms of the difference in the syllabuses\course types being followed in the two countries, thus introducing another confounding factor into the picture of the validation of oral rating scales, namely that of teaching programme. Ingram's study in this respect was flawed, as he did not control for or take into account previous exposure and
practice of the groups of students who were used in the validation study, as recommended by Hughes (1981: 177). Control for elicitation technique and scoring procedure between the CELT and the ASLPR was impossible, as the CELT does not contain an oral test.

Factor analytic studies have also dealt with direct oral tests in comparison to tests of other skills using pencil and paper elicitation techniques, providing evidence of divergent validity. Vollmer and Sang (1983: 62) show for example that FSI interview scores load on a different factor from a cloze test, the TOEFL battery, a test of listening, structure and reading. The same is true of the more recent concurrent validation study of the English Language Testing Service (ELTS), where the criteria chosen were the English Proficiency Test Battery (EPTB) and the English Language Battery (ELBA). Criper and Davies (1988) acknowledge that the choice of criterion is a problem for assessing convergent validity of the oral rating scale of the ELTS test, and indicate that although they had studied the relationship of the ELTS oral test to the oral component of the Association of Recognised English Language Schools oral test (ARELS) and the Test of English for Educational Purposes (TEEP), which are both SOPIs in that they are tape mediated, the relationship was "surprisingly weak" (Criper and Davies, 1988: 50). Unfortunately, Criper and Davies do not provide figures to illustrate this conclusion. Rea and Wall (1988: 58) reported correlations of ELTS - TEEP oral
components at .51 and ELTS - ARELS at .62 in a small scale study which they had conducted. James (1988) who developed the TEEP oral makes no comment on its relationship to other measures at all, either direct or semi-direct.

Clark (1975: 19) could legitimately claim that the research necessary to investigate the relationship between direct oral tests and indirect oral tests was "for the most part lacking." There was at least one exception to Clark's generalization. Carroll (1967) compared the MLA Proficiency Battery (1965) with FSI-type oral interviews in four languages, and produced correlation coefficients of between .66 and .82. A study by Clifford (1978) which also used an FSI-type oral interview and the MLA Cooperative Proficiency Test in German, producing a correlation coefficient of .83, was the only additional study which Clark could quote in 1979 (Clark, 1979: 40). Clifford's study was repeated by Dugan (1988) using the MLA English test and an ACTFL OPI, in which moderate correlations in the region of .7 were reported.

Cartier (1980: 8) reported correlations of .74 and .72 between the Defense Language Institute's Taped Oral Proficiency Test of Spanish (TOPT) and FSI interviews, and Lowe and Clifford (1980: 37) reported a correlation of .90 (using 27 students) in a comparison of the Recorded Oral Proficiency Examination (ROPE) developed for TESOL, and the Central Intelligence Agency adaption
of the FSI Oral Proficiency Interview.

By 1980 the reason for carrying out such concurrent studies had become plain: there were many situations in which there was neither the time nor the resources to conduct direct oral interviews, which in the studies mentioned above are considered to be superior to indirect oral tests. The motivation was one of practicality, the method was concurrent validation to decide whether or not an indirect test could legitimately substitute for a direct test, and hence strict control over facets which would allow the convergent or discriminant validity was not an important issue. Clark (1979: 48) for instance, states that:

"Semi-direct tests may be proposed as second-order substitutes for direct techniques when general proficiency measurement is at issue but it is not operationally possible to administer a direct test. In these instances, it is considered highly important to determine - through appropriate experimental means - a high level of correlation between the two types of instruments when used with representative examinee groups."

The position that semi-direct tests are "second-order substitutes" for direct oral interviews would seem to have changed very little since (Clark, 1988: 197, Clark and Lett, 1988: 13). despite considerable evidence that correlations above .80 can be achieved fairly systematically on well designed semi-direct tests of oral proficiency.
For example, the final version of the TSE was validated against the FSI in a concurrent correlational study (Clark and Swinton, 1980a), with correlations ranging from .73 to .77, as can be seen in Table 2.5 in the Appendix to Chapter 2. In the third edition of the TSE Manual for Score Users (1990) the two research reports of Clark and Swinton (1980a and 1980b), and the correlation matrix in Table 2.5 are still quoted as the only source of evidence for the validity of the TSE in the section dealing with the psychometric character of the test. The magnitude of these correlations has been interpreted as meaning that an OPI is superior to a semi-direct measure which can only be shown to have moderate correlations with scores from face-to-face interviews. However, Clark and Lett (1988: 13) report on the development of semi-direct tests which are correlated with the AEI, descendants of the FSI, in which correlations of .89 to .96 are quoted as validity coefficients. This refers to the work of Clark (1986b), Clark and Li (1986) and Clark (1988b) in the development of a semi-direct test of Chinese.

It would not seem unreasonable to comment that these studies beg the question of validity. It has been assumed that the FSI Oral Proficiency Interview is valid because of its "direct" nature (see pages 54 - 56), and hence safe to assume that reasonably high correlations between a semi-direct test and an OPI mean that the semi-direct test is valid. However, the driving force of practicality
has meant that the key issue of establishing the validity of one measure before others can be compared with it has been overlooked.

More recently, Davidson and Bachman (1990) looked into the relationship between the five FCE papers and the Test of English as a Foreign Language (TOEFL) as a part of a larger comparability study. In this case, the FCE oral test was compared with the rating components of the TSE in its institutional form, called the Speaking Proficiency English Assessment Kit (SPEAK), which is also a semi-direct tape-mediated test. They claimed to have found evidence of a distinctive oral trait. Abraham and Plakans (1988) reported a concurrent study between the SPEAK and Non-native speaker teaching assistant ability to speak English in the classroom at Iowa State University, and concluded that SPEAK was a good predictor.

The most recent work in concurrent validation has been on the development of SOPIs (Stansfield, 1989), the aim of which is to model face-to-face communication as closely as possible, pitched at the Intermediate, Advanced and Superior levels of the ACTFL Guidelines. Stansfield (1990a: 229) specifically refers to the work of Clark (1986b), Clark and Li (1986) and Clark (1988b) as the beginning of the development of the SOPI, which inspired the development of the Portuguese Speaking Test (PST), in which correlations between the SOPI and the OPI were found to be of the order of .93 (Stansfield, 1990b).
This led to the development of further SOPIs in other languages including the Hebrew Speaking Test (HeST) (Shohamy, Gordon, Kenyon and Stansfield, 1989) which produced correlations of .90 (Israeli version) and .94 (US Version) with the OPI (although Shohamy and Stansfield, 1990: 87 report these as .89 and .93 respectively, with a sample of only 20 students in each category, whilst Shohamy, Shmueli and Gordon (1991: 6) confirm the second set of correlations but claim the use of 40 students!), and Indonesian and Hausa (Stansfield and Kenyon, 1989). The Stansfield version of these correlations is reproduced in Table 2.6 in the Appendix to Chapter 2.

Stansfield (1990a: 232 - 233) argues that because the SOPI uses the format of an OPI, and similar tasks:

"...it seems time to reconsider Clark’s characterization of semidirect tests as "second order substitutes" for the direct OPI. While this characterization may be applicable to semidirect tests in general, it does not seem to apply to the SOPI."

Advantages of the SOPI over the OPI are stated to be increased reliability because of the removal of the interviewer factor, the standardization of delivery and content, and the increased amount of language sampled (ibid., 232; Shohamy and Stansfield, 1990: 81 - 82). However, Shohamy, Shmueli and Gordon (1991) do in fact report that despite the much longer time spent on doing
the SOPI there was no significant difference in the amount of language sampled when compared with an OPI.

Shohamy, Shmueli and Gordon (1991) argue that the correlation coefficients in Table 2.6 were not in themselves adequate to claim validity, and proceeded to examine the SOPI in comparison to the OPI in terms of functions and topics, linguistic errors and discourse features from test data, and finally features of orality. In terms of functions, it is argued that high level OPI students are asked to perform more functions than would be the case on a SOPI, but low and intermediate level students would be faced with significantly more functions on the SOPI than the OPI. The same is observed with variety of topic, with trivial topics being reserved for low level students on the OPI, whilst in the SOPI all students are exposed to the same topics. Shohamy et. al. see this as an advantage of the OPI, but this may not be necessarily true if task and topic affect anxiety of lower level students, as was found in this study (see page 256). It could well be that the non-adaptive form of the SOPI induces additional anxiety which would confound test scores, and this has not yet been investigated.

No significant differences were discovered between the two tests in terms of the number of linguistic errors made, although when turning to discourse features, self correction and paraphrasing were reported to occur more often in the SOPI whilst shift to L₁ was more frequent in the OPI. The explanation given for this by Shohamy et.
al. is that test takers view the SOPI as an elicitation procedure in which they are encouraged to concentrate on accurate production, whilst OPI test takers are involved in a communication activity and must focus on the production and negotiation of meaning. In terms of lexis, this latter feature has been observed before in the attempts of less able students to communicate (Fulcher, 1989), and the explanation is quite convincing.

With reference to recent debates regarding the cline between oracy and literacy, and the features of each which are manifested in different contexts (Halliday, 1985), Shohamy et. al. found that the SOPI contains fewer oral features than the OPI, in that the latter contains a higher ratio of grammatical items to lexical items, it is more contextualised and there is more personal involvement on the part of the test taker, including paralinguistic behaviour and the expression of personal views and emotions.

It would appear from this evidence that the OPI and the SOPI do in fact measure different abilities or traits. Why, then, did Shohamy et. al. (1991) find no significant difference in scores between the SOPI and OPI scores? The answer which they provide is that both of the tests were rated on the same ACTFL scale. If this is the reason, then this would provide evidence of convergent validity for the ACTFL scale as similar results are obtained irrespective of the data collection method. This deserves further investigation, not in the context of the
comparability of two sets of elicitation techniques, but in the context of convergent validity. Although convergent validity is not in itself sufficient evidence of the validity of a rating scale, it is necessary.

From the preceding discussion, it can be seen that the concurrent validation of two "direct" oral tests is rarely done. This in itself may very well be a significant observation: to conduct a concurrent study which would provide interpretable data would require that the measurement properties of the rating scales in the two direct methods were well known and capable of comparison. The selection of the criterion is therefore extremely important, and it is still true to say that "there is little agreed basis for comparing any two use-of-language tests" (Lee and Low, 1982, 16). Bachman and Palmer (1981a; 1981b; 1983) have shown how such comparisons might be made through Multitrait-Multimethod studies, but their work is concerned only with identifying an "oral trait" as opposed to a "reading trait" using only one direct measure.

One study which has investigated the use of a componential oral rating scale in a "direct" oral interview is that of Yorozuya and Oller (1980). Yorozuya and Oller avoided problems associated with rater training in that they used graduate students of linguistics who received no training or explanation of the scales used to rate students from audio cassette (ibid., 138). The scales used were marked Grammar, Vocabulary,
Pronunciation and Fluency, and consisted of ten "bands". Yorozuya and Oller concluded that there was no reliable specific variance for any of the rating scales, and that as a large g-factor accounted for the data a global rating scale should be preferred. However, the study is flawed in two ways which make the conclusions suspect. Firstly, the four rating scales used had no band descriptors: it was assumed that the component constructs would be understood by the graduate students from their label alone, and that each component would be construct valid if and only if it had a large degree of specific variance. This could have been just as much a test of the conceptual understanding of the raters as of the construct validity of the components. Secondly, the method used to analyze the data was Principal Components Analysis which reduces the data to the minimum number of factors possible (Woods, 1983), and given that only one factor was extracted we may suspect that the method could have imposed the solution on the data, given that there was no theoretical basis for expecting anything else to occur.

In Chapter 9 a concurrent study will be reported between the rating scales developed and described in Chapters 4 and 5 and the components of the First Certificate in English of the University of Cambridge Local Examinations Syndicate. The results of the concurrent study throw further light on issues in concurrent validation in oral testing outlined above.
2.2.3 Controversy over band descriptors in oral rating scales

Amongst the most vehement critics of the Proficiency Movement and the rating scales upon which it is based are Lantolf and Frawley (1985; 1988), who believe that the whole approach to oral testing is flawed.

With reference to the Proficiency Movement and the curriculum implications which are being implemented throughout the United States, Lantolf and Frawley point to a number of inconsistencies which cast doubt upon the whole process. They review two studies concerning the number of hours required to achieve the 2/2+ level. In Western Europe Brod (1982) claims that approximately 480 hours of study are needed to reach this level, whilst Lambert et. al. (1984) claim that some 840 hours of intensive study is required. Given such different claims, and unfortunately there is no published evidence to suggest that either claim is accurate, Lantolf and Frawley ask how it is possible to proceed with curriculum reorganisation based upon descriptors from AEI tests. From our concern with rating scale development, it may also be possible to interpret these disparate findings in the light of different researchers interpreting the rating scale in different ways and hence producing radically different estimates of the time it would take to reach a given level.

Referring to the topics chosen for the tasks, Lantolf and Frawley also question the assumption that it
is possible for students to achieve greater accuracy on topics which are known rather than those which are not known (ibid., 2 - 3), an explicit assumption of the rating scale band descriptors. Barnwell (1987: 40) makes the same point. The issue here is that no evidence has been offered to date which suggests that there is a relationship between accuracy of language production and the degree of "abstractness" or familiarity with the topic as posited in the rating scale. Douglas and Selinker (1985: 206) argue that the more familiar the topic the more likely the test is to "engage the test-taker's ability to perform." Working from within the field of discourse domains, they present the view that prototypical domains, such as "life-story" are more likely to engage the subject in this way. Fluency is seen to be a function of the student's ability to relate to the discourse domain chosen for the interview. This issue would become extremely important if future research were to discover that choice of topics does have a significant effect on the use of rating scales which are linguistic in nature, rather than referring specifically to given topics or situations.

Another assumption questioned by Lantolf and Frawley is that there is a difference between the ability of speakers used to dealing with foreigners and those who are not, in comprehending speech. This difference is built into the band descriptors in a number of places, but it is an assumption which has not been verified.
One of the few studies which has been used to support this distinction within the scales is that of Chastain (1980), which reported on reactions by native speakers to errors made in Spanish by intermediate students. The main problem with Chastain’s study is that the "errors" which were presented to the naive native speakers were constructed one sentence examples written on paper. Although Chastain acknowledges the limitations of the study and states that the results would need to be subject to replication (ibid., 214) he does not see any problem with the basic research methodology when it comes to applying the findings to naive native speaker reaction to speech samples. Van Patten (1986: 62) is particularly critical of this kind of generalization, stating that the lack of a database does not help the cause of the Proficiency Movement. One cannot generalize from one study such as that of Chastain, especially when other researchers using a similar methodology have come to the opposite conclusion, namely that native speakers are generally extremely tolerant of a student’s attempts to communicate compared to the harshness of non-native teachers familiar with the student’s L1 (Sheorey, 1986).

Barnwell (1989) has studied naive native speaker reaction to actual interview data, and found low rater-reliability and an inability to relate native speaker perceptions to the ACTFL/ETS rating scale without training. In this case the criticisms of Lantolf and Frawley have been upheld in the light of continuing
The main problem which Lantolf and Frawley see in the AEI rating scales is the underlying philosophical assumptions in the positions taken with regard to validity.

They claim that the approach to testing developed is "analytical" rather than empirical, but the analytical approach is presented as if it were empirical truth. The concept of the Functional Trisection is singled out for particularly harsh criticism. Lantolf and Frawley argue that the statements made by Liskin-Gasparro (1984b) and Lowe (1985a) "may have nothing to do with real world performance" and that "such criterion referenced tests impose competencies on examinees and measure the extent to which the person deals with the imposition" (Lantolf and Frawley, 1985: 339 - 340). They make a distinction between test language as reflected in the rating scales and real world language.

In order to grasp the arguments of Lantolf and Frawley a fairly long quotation is provided (ibid., 340):

"An informal look at the logic of speaker-levels and linguistic guidelines for determining performance underscores our argument: 1) assume, from informal observation, that some speakers of L2 are better than others, and rank these speakers in groups labelled from one to five (or one to three, in the case of the ACTFL/ETS Guidelines); 2) With speakers grouped, abstractly, from one to five, decide how to put people into these groups; 3) define each of these a-priori groups in terms of linguistic criteria also organised from simple to complex (i.e., 1 to 5). It is not
difficult to see the analytic logic here. The levels (1 to 5) do not exist except in terms of linguistic criteria which define them. That is, the logic of the levels and of their criteria is symmetric implication: $X \text{ (levels)} \iff Y \text{ (criteria)}$. What is a "low-novice?" Someone "unable to function in the spoken language." What is someone "unable to function in the spoken language?" A "low-novice." This logic does not yield criterion referenced tests, but criterion-reductive tests: the criteria are the levels and vice versa.

Because the levels are criterion-reductive, many of the statements made by researchers concerning the levels and the criteria are understandable, if erroneous. Liskin-Gasparro states that the Guidelines are "absolute definitions." They must be "absolute definitions" since they are analytic and reductive. The criteria, as absolutes, are then converted into requirements, because the criteria are required absolutely to define the levels. When the criteria are converted into requirements, one can argue, as Liskin-Gasparro and Omaggio do, that speakers must be judged in relation to their acquisition of the requirements: i.e., the acquisition of the analytically derived and absolute requirements which define each level - the acquisition of the so-called "competencies which lie at each level." With the criteria as analytic and reductive, it is impossible to evaluate L2 speakers with regard to the Guidelines. It is only possible to evaluate L2 speakers as they are in the Guidelines, because the Guidelines are absolute and reductive."

Lantolf and Frawley therefore claim that learners are being evaluated by a series of discrete stages through the law of non-contradiction: one cannot be in one level and in another. Level descriptors logically exclude all other levels. When it comes to applying these to the curriculum, Lantolf and Frawley argue that this amounts to saying that the "reductive definitions" cannot only be used for testing, but that they are also
teachable.

Further, because the definitions of the levels are measured against an "educated native speaker" they claim that far from being a criterion-referenced test, the level descriptors are in fact norm-referenced. Lantolf and Frawley question the ability of researchers to define an educated native speaker, but we will return to this later.

In their follow-up article, Lantolf and Frawley (1988) refer to the "psychometric posture" of the Proficiency Guidelines. They claim that the rating scale provides the "illusion" that something is being measured, whereas in reality no metric exists. They provide as evidence the fact that the FSI has five levels within its scale, the ASLPR has nine, and Bachman and Savignon (1986) suggest that there are really only three. Lantolf and Frawley agree with Vollmer (1983) that "proficiency" is what proficiency tests measure, and this will continue to be the case until a more empirical approach to rating scale development and number of levels described is attempted. There is a great deal of truth in this position, as Pollitt (1991: 90) would agree. Until it is possible to show that what is described in the bands of the rating scale relates to the way in which students actually do acquire language and how they really speak, rating scale constructors will be open to this charge (Pienemann, et. al., 1988).

The work of Lantolf and Frawley is timely, for the
issues which they raise are clearly important with regard to the development of oral tests and, specifically, rating scales. It cannot be denied that the AEI rating scales represent a-priori definitions of language proficiency which have not been validated prior to their use, and this has been a serious error on the part of the test developers. The fact that researchers have to rely on the defence of "experience" when facing criticism is evidence of this.

However, the work of Lantolf and Frawley must be viewed with extreme caution. Although their criticism of the AEI rating scales is largely justified, their solutions to problems are not. They argue that it is necessary to develop a learner-centred view of communication without the use of psychometric theory, and that the scoring procedure (no alternative is suggested) should represent "real-life" facets. How is this to be achieved? Certainly through the study of what learners actually say and the development of a database from which scales may be developed empirically (Fulcher, 1987), but this would imply the concurrent use of psychometric theory, which Lantolf and Frawley would outlaw. Indeed, it is difficult to understand the relationship between the criticism of the AEI rating procedure on the grounds of lack of an empirical basis and the suggestion that better scales could be generated without the use of the tools that make the empirical basis possible.

The contention here is that the criticisms of
Lantolf and Frawley are a remarkable indictment of the ACTFL/ETS rating scale, but that the conclusions are not valid. This may be seen most clearly in Lantolf's (1988) review of Underhill's (1987) work on oral testing, in which he writes:

"Underhill is to be applauded for his commitment to the humans involved in the testing process. Test designers have given privileged status to the testing instruments themselves and to the ubiquitous statistical procedures used to corroborate research hypotheses and have virtually ignored the individuals subjected to the imposition of a test (see Lantolf and Frawley, 1985, 1988). As the author cogently remarks: In a genuine oral test, this order of priorities is reversed. Real people meet face to face, and talk to each other...it is the people and what passes between them that are important, and the test instrument is secondary."

Lantolf believes that the work of Underhill (1987) is "a step in the right direction." This helps to clarify the suggested solutions in Lantolf and Frawley (1985; 1988). What they seem to be suggesting is a commitment to a human approach which relies upon common sense and the compassion of the tester/interviewer. It would appear that what is really being advocated is nothing less than the priority of "experience" over research, exactly what they have criticised the AEI test developers for.

Lantolf's review of Underhill (1987) should be compared with that of Fulcher (1990: 81) in which it is argued that
"...if this work is in any way a reflection of the "state of the art" in oral testing theory and practice, then our profession has developed little beyond the anecdotal state of presenting evidence for preferences and points of view."

Underhill (1987) does not contain any critical discussion of oral tests in common use, and the treatment of rating scales (ibid., 98 - 100) is limited to keeping rating scales as short as possible to avoid trouble in their construction, and to writing and improving band descriptors by "trial and error". The focus of Underhill's argument is that rating scales describe "typical students" and as there is no such thing as a typical student (ibid., 99) one must not expect rating scales to be as useful as the subjective judgement of the experienced rater. The brief treatment of "mark categories" (ibid., 95 - 97) includes concepts like "Hesitation (how much does the speaker hesitate before and while speaking?)" although there is no indication as to whether students would be expected to hesitate more at lower ability levels and less at higher ability levels (as assumed in most rating scales) or whether such a criterion might be influenced by other factors. Experience, and trial and error are advocated in place of research into the construction of rating scales and the rating process.

These reviews help to place the work of Lantolf and Frawley in an appropriate context. As a criticism of a particular oral testing scale their work was full of
insights and warnings (as Bachman and Savignon, 1986: 382 would agree). Unfortunately, they have generalized beyond the bounds of that particular rating scale and in their conclusions opened themselves up to serious criticism in turn.

So far, we have only considered the controversy generated by the work of Lantolf and Frawley on the AEI approach to testing, which is centred around the argument that the rating scales in the AEI tradition are circular and that there is no evidence which could support a claim to validity.

A critical approach to the ACTFL/ETS rating scale is also adopted by Bachman and Savignon (1986), but the nature of their observations is such that the relevance of the comments may be extended beyond the confines of the ACTFL/ETS to other oral tests and oral test development in general. As we have seen, one of the main defences of the ACTFL/ETS rating scale is that of "face validity" of the tasks and elicitation techniques, and "experience" in the use of the scale. Bachman and Savignon (1986: 381 - 382) argue that the fallacy lies in the notion that "direct" tests as opposed to "indirect" tests lead to the production of "normal" and "real" language. Direct tests have been considered to be automatically valid. We have seen that this is the position of FSI and AEI proponents (Wilds, 1979; Lowe, 1983; Lowe, 1986b; Lowe, 1987; Liskin-Gasparro, 1984a). Bachman and Savignon raise two objections to this
position. Firstly, we have no definition of what "normal" language is. The rating scale band descriptors certainly do not provide any such definitions. Secondly, the argument presented for validity using the claim of "directness" confuses "behavioural manifestations" with the "construct". Bachman and Savignon (1986: 382) say:

"As with all mental measures, language tests are indirect indicators of the underlying traits in which we are interested."

In this respect, all oral tests are indirect, and therefore it is illegitimate to rely on face validity. It is necessary to demonstrate construct validity for any rating scale which is developed. The position of Bachman and Savignon is therefore in keeping with the Standards for Educational and Psychological Testing (1985: 9 - 19): the problem with the ACTFL/ETS and other rating scales is the confounding of language ability with test method facets. Bands on the rating scale must be defined independently of method, which at the present time they are not. This is a clear argument in favour of research into the construct validity of rating scales independently of tasks and elicitation techniques.

Indeed, Bachman (1988) argues that in order to investigate the construct validity of oral rating scales, it is necessary to separate clearly the abilities to be measured from the tasks and elicitation techniques used. The problem is compounded where the single rating scale

86
developed represents a global assessment of oral ability, which has "no basis in either current linguistic theory or empirical work in language testing" (Bachman, 1988: 63). Bachman proposes the development of separate rating scales which represent a theory of the structure of oral abilities through his model of Communicative Language Ability (CLA), in order to distinguish these abilities from test method facets. These facets were first described by Carroll (1968) as: the testing environment, the nature of the test instructions, the nature of the input the test taker receives, the nature of the response to the input, and the interaction between the input and the response. Bachman (1990: 119) has expanded these in his categories of test method facets. In any study of construct validity, as a potentially large source of error, test method facets that can be identified should be controlled for or the extent of their effect on scores estimated, as argued by Shohamy (1988).

Bachman recommends the development of scales which are defined "abstractly" without reference to specific test method facets and the investigation into the way in which test method facets effect the scoring procedure. In the meantime, validation of the ACTFL/ETS and other rating scales remains impossible. In particular, the claim to having had years of experience working with the scale "...in no way constitutes evidence for validity" (Bachman, 1988: 163).

A response to Bachman came in a study by Dandononi
and Henning (1990) which used a Multitrait-Multimethod (MTMM) design to investigate claims that test method and trait were confused, and a Rasch analysis to investigate the extent to which the ACTFL band descriptors could be said to represent an acquisition hierarchy. The results of the MTMM study are reproduced in Table 2.7 and the Rasch analysis results for the oral test in Table 2.8, in the Appendix to Chapter 2.

In this study the two raters used were defined as the method facets for the skill areas of writing and speaking, and the "skill modalities" (reading, writing, speaking and listening) were defined as traits. Method facets in the reading and listening tests were item formats (Dandonoli and Henning, 1990: 13). On the basis of the evidence, it was concluded that convergent validity had been demonstrated for all skill modalities, and that speaking and reading demonstrated divergent validity. Secondly, it was claimed that the Rasch analysis showed that there was an "adequate progression" with the exception of Novice-High which was more difficult to attain than Intermediate-Low.

Whilst the Dandonoli and Henning study does go some way to answering the criticisms of Bachman and Savignon (1986) it should be observed that in the MTMM the only test method facet to be specified in their design for speaking was that of raters, and only two raters were used. Further, these two raters were trained and accredited ACTFL/ETS raters. It is thus quite possible
that the results obtained by Dandonoli and Henning are an artefact of the rigorous training of the raters used. Secondly, in the Rasch analysis of speaking ability no test method facets were specified at all. Thirdly, two ratings of the same performance cannot be said to count as separate methods. Although the results do appear to be impressive, we are not only faced with the possibility that they too are an artefact of rater training, but they could have been caused by test method facets such as the task or elicitation procedure.

It is not hard to level such criticisms at post-hoc attempts to validate a testing procedure and a rating scale after operationalization and the raters fully trained. What has happened is that it has been made much more difficult to isolate and control for or investigate the effects of the test method facets which are confounded with the oral trait.

2.2.4 Conclusions

The issue of validity in oral testing seems set to remain a contentious issue, especially between those writers who continue to maintain that "experience" and "common sense" is all that is required to ensure validity (Underhill, 1987; Lantolf, 1988; Morrow, 1990) and those who desire to establish an empirical basis for the study of validity. Two key aspects of the validity debate in oral testing are important for the present study.

Firstly, it is important to design studies which
account for as many of the test method facets as possible. Little is known about the size of the effect of these test method facets, and it is only through research that they can be isolated and estimated, so that we can state whether or not a trait has been identified. We note that in MTMM studies to date, an oral trait does legitimately seem to have been identified, although no work has been done into the components of that oral trait: no validity study has been conducted on the Functional Trisection, on the FSI rating scale components other than Adams' (1980) factor analysis, and little convincing evidence is available to suggest that the band descriptors in rating scales represent external linguistic reality. Secondly, within a research framework which seeks to investigate the validity of oral rating scales it is of vital importance that raters are not trained, socialised or cloned, as this will inevitably have the effect not merely of confounding, but "fusing" together a test method facet with a hypothesized trait in such a way that they can no longer be separated. The investigation of validity then becomes impossible. Rater training should only be introduced at a later stage prior to the operationalization of the test, in order to ensure valid assessment in live testing.

The approach adopted in this research is to develop rating scales and band descriptors on the basis of an analysis of actual student speech (Fulcher, 1987; 1989). Aspects of student speech which appear to affect the
scores which students are awarded will be isolated and, through the use of discriminant analysis an attempt will be made to construct two rating scales. When these two rating scales have been used in trial testing a validity study, taking into account the two key issues to have emerged from the validity debate described above, should then be able to avoid many of the problems which have plagued oral testing research to date.

Section 3: Rating Scale Construction in Oral Testing

3.1 Approaches to constructing oral rating scales

We have looked at the most influential of oral rating scales in Section A, and considered issues of reliability and validity in Section B. However, we have not so far considered how these scales are actually produced.

Very little has been written on the actual process of scale construction for scoring oral ability, although as we have seen claims are made for the scales once they exist. We may justly say that the development of scales to date has been carried out, on the whole, using a-priori decisions regarding what the developers believe to be aspects of language proficiency which define a pre-set number of bands within a scale, based on experience of working with second language learners. In this process it is not uncommon for band descriptors within scales to be subjected to scrutiny by professionals within the field, but further than this there is no check upon the number
of bands or the actual content. Most claims are made for the band scales only after they have come into use.

Frith (1979: vi) acknowledges that for the FSI scales there are no external criteria against which they can be assessed; rather each level within a scale is defined in relation to the other levels and the only key reference point "the ultimate standard, the ultimate criterion reference, is the proficiency of the educated native speaker."

The emphasis in scale development has therefore not been upon the construction of the scales themselves, but achieving and demonstrating reliability of the scales once they have been developed. It may be suggested that this is a major potential source of invalidity within the testing procedure. Although Wilds (1979: 1) does claim that "the usefulness of the system is based on careful and detailed definition in both linguistic and functional terms of each point on the scale" she also goes on to state that the band definitions need to be "gross" in order to achieve reliability (Wilds, 1979: 2).

Jones (1975: 3) argues with reference to the FSI that

"One of the principal problems we are faced with is the construction of proficiency tests which measure language ability accurately enough to correspond to these definitions."

This shifts the emphasis away from the band
descriptors to the elicitation techniques. However, Jones’ point leads to a further observation which we might make: the definitions or band descriptors are assumed by Jones to have been written without reference to any specific testing procedure (or indeed to a database of L2 production), and so the problem which is then faced by the scale constructor is to develop a testing procedure which would allow scoring to be carried out on the scale which has been constructed. We have seen that this is not always the case, as rating scales do confound non-linguistic (situation or topic specific) and linguistic criteria. A similar point is made by Alderson (1991a), where it is reported that draft IELTS (International English Language Testing Service) band descriptors contained descriptions of performance which were not elicited by the actual tasks in the test.

Jones (1975: 4) further says that very little is known about tests such as the FSI, as "no validation studies have been made with the definitions as the basis". Jones (1981) makes similar points, adding that most oral testing procedures do not relate the elicitation technique to the scoring systems in any specific way. No recommendations are made regarding how to overcome this problem, although Jones appears to argue that a balance must be maintained between the assessment of overall competence and components of competence (such as fluency, accuracy etc.). Further, despite his concern with the association between band descriptors and
eliciting appropriate language, Jones does say that impressionistic scoring is to be preferred to analytic scoring. This inconsistency may have led him not to pursue the issue of scale development very far. Indeed, the general view that an "intuitive" approach to scoring is adequate is widespread. Berkoff (1985: 96) supports Jones' approach explicitly, claiming that analytic scales are "tedious" to use, and that the intuitive judgement of an experienced oral assessor is to be preferred. This is an argument in favour of abandoning rating scales as such, except in so far as they offer some reference point for the "number" given to the student and the training of the raters, so that high reliability is achieved by ensuring that all raters see the world through the same spectacles. Wilkinson (1968: 121 - 122) also states that while impressionistic marking is advantageous, the training of raters is essential if they are to grade in the same way. Lowe (1985a: 16) argues in the same vein and Underhill (1987) is a more recent virulent spokesperson for this particular approach to oral testing. This approach thus emphasises the importance of rater training at the expense of scale development and scale validation, as was argued previously (see page 49).

With regard to the ASLPR Ingram (1982: 9) says that in the scales "each definition exists in the context of the whole scale and in relation to adjacent definitions" whilst not describing how the descriptors for each band were originally written. He claims construct validity for
the rating scale merely by stating that when they were being produced the developers drew upon "psycholinguistic research" (ibid., 19). However, he goes on to claim that the scale is not impressionistic, as it relies implicitly upon the concept of a "universal pattern of language development" which is not described (ibid., 10).

Similarly, Ingram (1985a: 4) makes a stronger version of this claim when he writes:

"We have noted that rating scales such as the ASLPR are developmental in structure, i.e., the progression of behavioural descriptions attempts to reflect the way in which a second language develops from zero to native-like. Thus the scale is not arbitrary but related to the universal developmental schedule and a learner's rating indicates the stage of development he has reached in the developmental schedule."

This has come under criticism from Pienemann et. al (1985) because the notion of "development" must be theoretically coherent and empirically verifiable. The ASLPR does not seem to be empirically verifiable for Pienemann. Further, the definitions of levels arrived at on the basis of the developmental process must be quantifiable, or validation studies cannot be carried out. It would therefore seem that the ASLPR scales suffer from the same flaw as many other rating scales. This should not be surprising, in the light of the fact that the ASLPR was generated from early drafts of the ACTFL/ETS rating scale.
It would seem that these approaches to rating scale development place an unvalidated view of the linguistic world onto the raters and students alike. Students must perform in such a way that they can satisfy a scale which "intuitively" and inexplicitly reflects the subjective experience of the scale developer and, if it is "gross" enough, the subjective interpretation of the rater.

A far more satisfying approach is one where the band descriptors within the scales are developed through an empirically verifiable procedure, and

"will have to be based on observed learner behaviour as opposed to postulated or normative notions of how learners ought to behave." (Pienemann et. al., 1985: 2 - 3).

Observed learner behaviour must be quantifiable, and procedures developed for using the information in the construction of rating scales which reflect the actual linguistic behaviour of students. In this way the close relationship between linguistic behaviour (which may or may not be specific to the elicitation technique, task or topic), the elicitation technique, the task and the scoring procedure could be ensured. This may be characterised as a "data-driven" approach to scale development with a strong theoretical and empirical underpinning, rather than the generally accepted current approach of producing scales on the basis of a-priori decisions concerning the structure of proficiency. This
is an argument which would be supported by Bachman (1988; 1990). The main problem which may have to be faced in a data-driven approach is whether or not the scoring procedures could then be legitimately used in a range of testing situations other than those on which the scale was originally developed. This is an empirical question which can be addressed through using a rating scale over a number of different testing situations and assessing the extent to which scores on the scale differ across those situations.

Lowe (1985a: 19 - 25) goes some way towards describing the development of the ILR scales. The original FSI was a semantic differential with the intermediate levels representing a relative amount of a quality between bipolar terms. Later these were supplemented by verbal descriptions of the intervening points, creating a Likert scale. However, the scales were certainly a-priori, and the descriptors altered in the light of changing Second Language Acquisition (SLA) theory and experience of their use, not on the grounds of empirical research. As Lowe says, there is a "non-compensatory core" of language abilities which the students must possess for a given level within the ILR scale (i.e., without language ability X you cannot be placed in, say, band 2), whilst other aspects are compensatory (i.e., if a student does not possess language ability Y, s/he may still get into band 2 if s/he possesses more of language ability Z for the level).
As a student's level of proficiency increases, the more non-compensatory the scale becomes (ibid., 25). The descriptions of both compensatory and non-compensatory elements can change over time. Fluency, for example, has been "redefined" from "ease of speech" to include "the specific requirement for speech rate to fall within the range of rates used by native speakers" (ibid., 26). Statements such as these may give the impression that scale development and growth is empirical, but it is in fact still a-priori, as the "speech rates" have not been empirically established and, even if they were, they would still have to be related to all the test method facets in the data collection technique. This once again highlights the relationship between testing situation and the scoring procedures which any a-priori system cannot deal with adequately.

It is because of the a-priori approach to scale development that is currently in widespread use that "proficiency" is seen to be what proficiency tests measure (Vollmer, 1983), and that Lantolf and Frawley (1988) can argue that there is no evidence to suggest that proficiency is scalable. One symptom of this is the fact that no justification of the number of levels within a scale can be provided. By using observation of actual L₂ production on language tasks and the production of scales using discriminant analysis, these problems may be overcome, for the number of bands within a scale will be determined by the highest discrimination index between
students provided by the dependent variables in the analysis.

The resulting scale would be criterion referenced in that candidates would be graded according to criteria isolated from the observation of actual performance, although the actual criteria would, it must be admitted, be dependent upon the sample from which the criteria were developed. Whether or not this would turn out to be a problem in terms of generalizability would therefore become an area for investigation. However, Bachman and Savignon’s requirement that a criterion-referenced scale must have a zero and a perfect point on an absolute scale (Bachman and Savignon, 1983: 382) could not be met, and there is no reason why it should be. As Bachman and Savignon point out themselves, we do not test students with no language ability at all, and "perfect" language ability is a chimera. Rather, the criteria are drawn from actual use and thus, to some extent, rely on the "norm" for the sample of the population used to develop the criteria. However, it may be legitimate to include a band at the lower and higher end of the scale which have not been generated by the discriminant analysis. For example, in the Fluency scale which is described in Chapter 4 there are five bands. It is possible to add band 0, which is defined as any performance which is not good enough to meet the criteria for band 1, and band 6 which would be defined as any performance which is superior to that in band 5.
3.2 Psychometric considerations

Clark and Lett (1988: 20) summarise the psychometric concerns which must accompany scale development:

"In order for a system of ratings or descriptions to be considered a scale, the ratings or descriptions must (1) denote the relative presence or absence of some substance or quality and (2) be capable of validly and reliably ordering persons or objects according to the extent to which they possess the substance or quality at issue."

From this follow two important assumptions: firstly the quality which the scale developer wishes to measure must be scalable. Secondly, the construct which underlies the scale must be unidimensional. The a-priori approach to scale development attempts to achieve these qualities at the expense of producing scales which suffer from being absolute and hence circular (Lantolf and Frawley, 1985; 1988). Scales which are arrived at through discriminant analysis will help to ensure scalability during the actual process of producing the band descriptors and should give some confidence in psychological unidimensionality, which at present can only be assessed theoretically in accordance with the definition of the trait underlying the rating scale and its band descriptors (Henning, 1992: 11). Psychometric unidimensionality could only be investigated in the trialling of the rating scales themselves.

The approach to scale development which is being
suggested here redresses the imbalance between the use of linguistic criteria and psychometric criteria (Perren, 1968: 109) by using speech as the basis for the development of the scales. Psychometric tools aid the process of scale development, but do not entirely dictate their format or content. The enterprise is thus a combination of linguistic analysis and interpretation, and psychometric rigour.

3.3 Terminology used in constructing band descriptors

The a-priori approach to scale development described above has led to a certain amount of vagueness and generality in the descriptors used to define bands. Of course, for some writers this is seen to be a strength (Wilds, 1979; Bachman, 1990: 341 - 346), as the generality of the scale and the gross nature of the bands mean that it can be applied to the assessment of any language in any testing situation. The disadvantages are, however, much greater than the advantages. Lack of clarity and vague use of key words within definitions detracts from the independent meaningfulness of the band descriptors. Investigation of the validity of the scale can only be conducted on the basis of test scores which may be an artificial product of rater training rather than valid measurement of qualities of the scale.

Schultz (1986: 373) claims that the ACTFL/ETS Guidelines provide testers and curriculum designers with a common terminology within which they can work for the
first time. This optimistic view reflects the claims of Wilds (1979: 1) that key terms such as "good" and "fluent" used prior to the FSI descriptors were extremely vague and, being open to as many interpretations as there were raters, were in need of revision. Wilds’ view is that the FSI descriptors are detailed enough to remove the uncertainty of interpretation which bald descriptions create. Wilds’ view is echoed by Lowe (1983: 231) in relation to the ILR rating scale.

Typical band descriptors are, however, far from self-explanatory (Barnwell, 1989), and as Matthews (1990: 119) notes, the bands

"are described in only vague and general terms and abound in qualifiers, so that only gross distinctions can be made with any confidence."

Alderson (1991a) testifies to the problem faced when constructing band descriptors, but argues that what matters at the end of the day is that assessors agree on their interpretation of the scale.

In an analysis of references to vocabulary knowledge in the ACTFL scale, Fulcher (1989) concluded that the descriptors were far too vague to be of any practical use within an operational oral testing procedure because they were not based on actual language production. Matthews (1990: 119) and Fulcher (1987) both draw attention to unclear statements regarding fluency in the ELTS rating scale as being unhelpful to raters.
Hieke (1985: 135) takes a particularly adamant stand on this issue:

"A glance at the literature on fluency reveals it to be replete with vacuous definitions, overlapping terminology, and impractical assessment strategies."

Mullen (1978a: 33) for example, considers fluency the "easiest" aspect of oral assessment, as one need only consider pauses and fillers. Hieke (1985: 137) is probably correct when he argues that tests cannot be "fair" (which could be interpreted as "open to empirical investigation")

"as long as they (the descriptors) hinge upon prose statements to delineate levels while these are peppered with notions that cannot withstand close scrutiny."

Hieke's own suggestion is to develop a semi-direct test in which the candidate speaks into a tape recorder. The tape would then be used for grading fluency along the parameters: speech rate (syllables per second), length of runs (average number of syllables between pauses), and hesitation phenomena (stalls, repairs, parenthetical remarks, silent pauses, fillers, progressive repeats and false starts). This would be quantifiable, but the sheer increase in marking time per candidate over an oral interview would, despite Hieke's protestations, probably make the suggestion impractical unless it could be made
automatic (on the possibility of this see Pendergast, 1985).

Finally, Hieke (ibid., 136) makes a point which has not been made elsewhere. Whilst the terminology of the band descriptors is vague, it may be meaningful for experienced evaluators who have not been trained and, hence, socialised in the use of the scale. However, Hieke would still challenge any rater to say what this would mean for a given student in practice at any given band. That is, once again, the issue of rater training is pushed to the fore as the most important reason for rating scales appearing to be meaningful and providing reliable results.

Alderson (1991a) distinguishes between three uses for scales: for reporting student abilities to test users, for "guiding the rating process", and for "guiding the construction of tests." In the preceding discussion it is clear that we are primarily interested in the second of these uses for scales, which Alderson deals with in relation to the IELTS rating scale. It is pointed out that raters must rate to a common standard - which no one would wish to disagree with - but in order to achieve this, emphasis is placed once again on rater training. It is, in passing, noted that it is important to "produce Band Scales that are more easily interpreted and used...." The question that remains is to what extent the scales are valid in isolation from the training which raters must receive in order to become certified raters.
Training in this way is designed to mask problems with the wording of bands in the scale by creating the illusion of psychological reality through rater reliability.

However, the way in which the IELTS band descriptors were produced does appear to be something of an improvement upon other a-priori methods of development. Three methods were employed: (1) draft descriptions were sent to "moderators, teachers, experienced markers, and others" for comment, the results of which were used to improve band descriptors; (2) experienced markers were asked to match speaking samples with band levels and "key features" of the samples were then identified; (3) the process of scale use in assessment and feedback from raters provides a continuous supply of information which could be used to improve the wording of the scales.

In this process it may be suggested that it is extremely difficult to devise a way in which the rating scales themselves may be subjected to a study of their validity in isolation from the raters who, to some extent, already share a standard interpretation of a band by its number, irrespective of the descriptor given to that band. The principle of socialisation in common interpretation will influence all results and mask details of scale operation in which the researcher may be interested.

Before it is possible to achieve agreement through shared perceptions of ability by training, it is
essential to establish that the view of proficiency or language development contained in the bands of the rating scale is valid independently of trained raters. Only in this way can training, rating and results be investigated in terms of validity considerations. This argument may be supported by the fact, as reported by Alderson, that the IELTS revision project had to retain nine bands in the speaking scale because of the "intuitive understanding" which had been developed for each band. This requirement is irrespective of whether a nine band scale may be empirically justified. Alderson's distinction between the reporting function of the scales (reporting scores to test users) and the grading function of scales (making the task of the rater meaningful) provides the answer to this problem; the IELTS draft band descriptors were indeed produced for test users and not raters. The "dubious manner in which the original Band Scores were derived" has imposed an unvalidated view of language proficiency and development onto the tester whether he wishes it or not, and the function of this view of language proficiency is primarily to provide meaningful interpretation of test scores by test users who "have a feel" for the meaning of the bands. It would therefore appear that the function of reporting test scores rather than rating student performance has provided the main motivation in the development of the new speaking scale in the IELTS test.

This example from the IELTS serves to emphasise the
point already made, that the meaning of band descriptors in most current rating scales tends to depend on the "intuitive feeling" of the rater (or score user) once socialised into the use of the scale. This will tend to be inevitable as long as the scales are constructed on a-priori grounds, whether it be the IELTS, ACTFL, ILR or some other derivative such as the ASLPR (Ingram and Wylie, 1985) or the AIMS test (Reschke, 1984).

3.4 A special case of terminological confusion in oral rating scales: the Native Speaker Yardstick.

In 1968 Perren noted that "native" or "native like" language ability was a term which appeared in the top band of rating scales. "What does this mean?" he asked. "What kind of native, speaking about what, and to whom?" (Perren, 1968: 111). Perren recommended the development of scales based upon observations of the proficient second language speaker and not some notion of a "native" speaker - but until recently his views have not received the attention they deserve.

A-priori scale development generally has depended upon the concept of the native speaker or "educated native speaker" for the definition of the top band in a scale. Wilds (1975: 36) argued that the linguistic ability of a "well-educated native speaker" was the "absolute standard" upon which the FSI scales rested. As we have seen, the definitions of other levels were then hung upon this one peg through the principle of internal
Liskin-Gasparro (1984a) argues that the top of the ACTFL scale, representing educated native speaker abilities, is often achieved by students who have the opportunity to live in the country where the target language is spoken. Lowe (1983: 231) characterises the ILR oral assessment scale as ranging from "no practical ability" to "well-educated native speaker ability". One of the test's strengths for Lowe is that it judges the students' production in relation to "language as it is spoken by well-educated native speakers." It is not surprising to find that the ASLPR also depends upon the "native speaker yardstick" (Ingram, 1982: 8), and that the top band in the speaking scale represents "native-like" proficiency (Ingram, 1985a: 4; 1985b: 222 - 225).

The use of the concept of the educated native speaker is increasingly coming under attack, and with its demise one of the main pillars of many rating scales currently in use is removed.

The foremost problem with using the concept of the native speaker as a criterion, and thus as a peg on which to hang other bands in rating scales, is that native speakers "show considerable variation in ability" (Bachman and Savignon, 1986: 383). In their criticisms of the ACTFL scale, Bachman and Savignon (ibid., 385) point out that although Lowe claims to have abandoned the educated native speaker standard,
"the ACTFL scale definitions are firmly rooted in the misconception that we can clearly identify native speakers and their standard of language performance."

This can be seen in phrases from the band descriptors such as "can be understood by native interlocutors", and "...using native like discourse..." Bachman and Savignon conclude that the scale still contains "the notion of a monolithic group of native speakers...."

Lantolf and Frawley (1985: 343) also argue that current scales contain "an implicit notion of the mean linguistic behaviour of an ideal speaker." The problem for them is similar to that of Bachman and Savignon: we do not know what an educated native speaker is. Lantolf and Frawley correctly state that when using the educated native speaker as a criterion, scale developers are concerned only with "THE" native speaker, whereas in reality only types of native speakers exist. Lantolf and Frawley (ibid., 343) identify four "types":

1. Idiolectal, or informants.
2. Statistical, or typical speakers.
3. Normative, or expert speakers.
4. Former, or speakers from historical records.

ACTFL attempts to combine statistical and normative on an intuitive basis. We may add to the criticism by
noting that even "expert speakers" may only be "expert" within certain contexts, and not others. Jarvis (1986: 20) provides the example of a professor whose "competence" (we might wish to say "performance") varies depending on whether s/he is talking about an area of speciality, socialising at a cocktail party, or opening a bank account. Whatever the case, the ACTFL scales combine the statistical and the normative speakers to form the band descriptors through intuition and not empirical evidence. Hence, Barnwell (1987: 39) can legitimately claim that what is being dealt with is an "ideal" native speaker, not real native speakers. Lantolf and Frawley are correct when they say that as the concept of the native speaker or educated native speaker "is neither unitary nor reliable" it should be abandoned (or at least more thoroughly investigated) as a criterion. Davies (1990: 52; 1991) also demonstrates that the concept of the "native speaker" is one which is not consistent, and no researchers have sufficiently defined the term to make it useful in a testing context, and the results of studies generated by interest in "World Englishes" also seriously cast doubt upon the notion of the "educated native speaker" as construed by test developers (Kachru, 1992: 4 - 5). We must conclude that Perren's (1968) call for studies based on real second language learner speech should be heeded.

In the face of these criticisms, Lowe (1985a: 47; 1987) has stated that native speaker performance on AEI

110
tasks would in fact probably fall in band 3, or "Advanced". "Superior" performance is equivalent to educated native speaker performance, which most native speakers cannot achieve. No evidence is presented to support these claims, other than the usual appeal to intuition and experience that there is some coherence to the concept of the educated native speaker amongst "adepts".

The concept of the educated native speaker should be abandoned in rating scale development at least until such a time as it is more thoroughly investigated and understood - if this is indeed possible - and an attempt made to use the production of second language speakers to create the criteria to be used in band descriptors. This will provide the possibility of external validation for the rating scale. The a-priori method of scale development has no doubt been useful in the development of practical tests of oral ability which have been used with some degree of success in the rating of students. However, given the problems with a-priori rating scale development discussed above, it is an appropriate time to try the new approach of data-driven scale development. Potential advantages and problems of the data-driven approach can then be assessed. Those scales which are termed "Real-Life" scales by Bachman, such as the ILR, are primarily concerned with face validity and predictive validity: the reactions of the test takers to the test, and the assessment of the degree of success which might
be achieved in future performance in some non-test situation (Bachman, 1990: 301 - 302). As we have seen, content validity is claimed on the grounds that the test tasks reflect the "real-life" situation to which the test is intended to predict. However, on this basis it is not possible to investigate construct validity because the performance on the task is equivalent to the ability being measured, and test method facets thus confounded with traits. This may be perfectly justifiable if, but only if, the results of the test are not generalized outside the specific context of the test situation. In order to generalize, it is essential to be able to claim that the test is tapping some underlying "ability" which may manifest itself in a variety of circumstances, both in terms of test and non-test situations where language is used. Bachman (ibid., 341) suggests one alternative: the development of "absolute scales" of proficiency, by which he means an a-priori scale which has a perfect and zero point, irrespective of whether or not actual learners fall into these bands (ibid., 346), and which are not related to any specific language users (ibid., 344). They will differ from current scales of the AEI type in that they will not contain references to aspects of the situation of testing and the elicitation techniques used which are considered to be test method facets, but only the abilities to be measured (ibid., 347). This approach is still a-priori, and would result in scales which shared the feature of "gross" definition.
with the AEI rating scales, which Bachman criticises. This would mean that even if it were possible for a researcher to show satisfactorily that the scales developed were testing "ability", it may not be possible to make any meaningful statements about the ability in question. The key feature of a scale which is to be developed on data-driven principles is that the descriptors will be specific, and directly relatable to aspects of student performance in which the tester or researcher is interested. This is to say that whilst the notion of the "absolute scale" as presented by Bachman may allow for studies of validity, it would be trivial in the sense that the level of delicacy of analysis would be too gross to make any specific, interesting claims for the scales under development. However, scales which are specific allow for non-trivial claims regarding their validity with regard to the underlying abilities which they claim to tap, if results can be shown not to have been contaminated by test method facets such as task or elicitation procedure.
Chapter Three

Rationale, time scale and design of the study

Section 1: Rationale

It has been argued in Chapter Two that a-priori scale development is flawed, and that post-hoc validation of such scales is plagued with potential problems. It has been suggested that one way to tackle the problem of the validation of oral rating scales is to develop band descriptors from a database of what students actually say (Fulcher, 1987; 1989), and this argument seems to have been well-received (Shohamy, 1991: 125). The key rationale of the present study was to investigate the potential of a databased, data-driven approach to oral rating scale construction based upon as complete a study as possible of the operation of those scales in terms of reliability and validity. To this end, test method facets which have been identified as confounding reliability and validity studies were either controlled for (interviewer, rater training, place of test, variation in elicitation techniques, time) or estimated (rater, task, affective factors).

The main aim of the study is therefore to investigate the operation of rating scales produced using an analysis of a database of spoken English to see if any claim to their superiority over a-priori rating scales can be made.

The two rating scales developed for this purpose were an Accuracy rating scale and a Fluency rating scale.
The ELTS global rating scale was also included in the study to allow comparison between the two databased scales and one a-priori scale. The constructs of accuracy and fluency were chosen because they are basic concepts in all rating scales used in oral testing, whether the rating scales are holistic or componential (see Chapter Two, and reviews of other scales in Chapters Four and Five). It is widely assumed in oral testing that the two constructs of accuracy and fluency are separate aspects of oral ability (Griffin, 1985), and in language teaching it is common for a strong difference to be drawn between accuracy activities and fluency activities (see for example, Brown and Yule, 1983: 104; Brumfit, 1984; Rixon, 1992: 81). In the research literature this difference is variously referred to as that between "norm-oriented" and "communicative-oriented" learners (Clahsen, 1985), or "rule-formers" and "data-gatherers" (Hatch, 1974), or "planners" and "correctors" (Seliger, 1980). In each case, the former of the two categories refers to students who concentrate on building up grammatical rules and aim for accuracy in production (at the expense of fluency), and the latter category refers to students who concentrate on communicating fluently, paying little attention to accuracy or alternatively, using their knowledge of the grammar and lexicon of the language only to correct themselves after they have communicated. This "basic polarity" as Brumfit (1984: 50 - 57) describes it, predicts that these two components of oral proficiency
will tend to develop separately depending upon the learning orientation of the student. Such a difference would also seem to be predicted by the distinction often drawn between "gestalt" and "analytic" learners (Dulay, Burt and Krashen, 1982: 237 - 238; Larsen-Freeman and Long, 1991: 196 - 197), and the two-dimensional modular theory of second language acquisition which posits knowledge-based and control-based language processing (Bialystok and Sharwood Smith, 1985).

Although little empirical research has been conducted into the distinction between students who are "norm-oriented" and those who are "communicative-oriented" to date, Ellis (1990a) conducted a study in which it was hypothesized that students who focus on accuracy would acquire linguistic knowledge quickly, whilst students who focus on fluency would develop "channel control mechanisms" (measured as speech rate) much more quickly. Using correlation and a Principal Components Analysis, Ellis (ibid., 89 - 90) concluded that

"...those learners who showed the greatest gain in acquisition of the three word order rules manifested the smallest gain in general oral fluency and, conversely, those learners who developed the ability to process speech the most rapidly displayed the smallest gain in knowledge of the word order rules."

As this distinction is an area of current research in Second Language Acquisition, researchers do seem to be
agreed that the distinction is a real one, the distinction is common in most oral rating scales currently used, and a considerable amount of research into the concepts of accuracy and fluency has been conducted (see Chapters 4 and 5) which has also provided methods for analysis and categorization, rating scales of Accuracy and Fluency were developed for this study with the hypothesis that the two constructs would provide a reasonable opportunity of demonstrating divergent validity.

Despite the reasons provided for choosing the constructs of accuracy and fluency described above, the labels "Accuracy" and "Fluency" should be interpreted by the reader in the light of the definitions provided in Chapters three and four, as they may not always correspond exactly with definitions or descriptions found in some of the literature.

Other available componential models of ability developed for language teaching and testing (Canale and Swain, 1980; Canale, 1981; Canale, 1983; Bachman, 1988; 1990) were thus not used, as the extent of their validity as models has hardly begun to be investigated. Further, for a databased study of this size it was considered preferable to work with a very limited number of components, the definitions of which are less problematic at the operational level than, for example, "illocutionary competence" or "sociolinguistic competence" (Bachman, 1990: 87), although Bachman’s
"Grammatical Competence" which includes lexical competence, is very similar to what is defined in this study as "accuracy".

Section 2: Time Scale and design

The spoken data used for the development of the two oral rating scales described in Chapters 4 and 5 was collected in 1989. Twenty one students took the ELTS speaking test, which was recorded with the permission of Nick Butler, then ELTS officer at the British Council, London. The tests were held on three separate days within the space of one and a half weeks. The data was collected using a miniature audio recorder and a tie microphone, and transcribed for the analysis. All students were native speakers of Greek.

Transcripts were made from the audio tapes, and the transcripts coded for aspects of disfluency and errors in grammatical accuracy as described in the literature and outlined in Chapters 4 and 5. This coding was done using the transcripts, the recordings, and the interviewer's recollections of the students from the actual interviews. This necessarily involved a degree of judgement on the part of the researcher, and sample transcripts are included in the Appendix to Chapters 4 and 5 for the reader to examine.

The twenty one students in the sample were grouped into ELTS bands, and the tallies of disfluencies and errors were summed for each of the groups. Group means
were calculated for each category of analysis within the definition of Fluency and Accuracy, and discriminant analysis used to discover whether or not these categories were capable of discriminating between the groups of students. When this proved to be successful, band descriptors were written using the original category descriptions and the actual data from the interviews.

Two rating scales were generated in this way, one for the measurement of fluency and one for the measurement of accuracy, as described in Chapters 4 and 5.

In order to study the reliability and validity of these two rating scales the three tasks used in the research were developed in 1990. They were informally pre-tested on a group of five students and then slightly adjusted to take into account potential problems in the smooth transition between the Warm-Up, Task and Wind-Down phases. The tasks were conducted in February and March 1991, with 47 students of English in Cyprus. The procedure used in the conduct of the tasks is described in full in Chapter 6. All interviews were recorded on video for future rating.

Each student took the three tasks within no more than one and a half weeks, and immediately upon completion was asked to fill in an extensive questionnaire on their test taking experience. The analysis of the questionnaire results formed the basis for the investigation of affective factors, which were
considered to be one facet of the test method. The interpretation of the questionnaire is reported in Chapter 7. Eight of the 47 students took part in a retrospective study two weeks after they had completed the oral tasks. It was hypothesized that such a study would throw light on test method facets of which the students were consciously aware when taking oral tests, and that the data gathered would provide extra information which would feed into the validity study of the rating scales by indicating whether or not aspects of facets other than trait were systematically affecting scores. The observations and opinions of each student were recorded on audio tape after they were shown an extract of one of their tasks on video. The transcripts of these interviews were used as the basis for a retrospective study reported in Chapter 10.

Five raters were asked to rate all 47 students on the three tasks on the Fluency rating scale, the Accuracy rating scale and the ELTS global rating scale as quickly as possible during the summer of 1991. Copies of the video cassettes were given to the raters at the beginning of August 1991, and were returned by mid-September. Each rater therefore had to award a total of 423 scores, providing 2115 separate ratings. These ratings were used as the basis for the reliability and validity studies of Chapters 8 and 9. Reliability was investigated using classical analysis, a G-study and a Rasch Partial Credit model. Concurrent validity was studied using student
scores on the University of Cambridge First Certificate in English and concurrent teacher assessment, employing correlations and Principal Axis Factor Analysis. Construct validity was investigated using a Multitrait-Multimethod design, Maximum Likelihood Confirmatory Factor Analysis, and a Rasch Partial Credit Model.

The raters were also asked to keep notes on any problems observed or difficulties they had while rating, and each of the raters was then interviewed. The interviews were recorded on audio tape and later transcribed. The retrospective rater study is described in Chapter 10.

The analysis of the data was carried out between September 1991 and May 1992.
Chapter Four

The development of a Fluency rating scale

Section 1: Introduction

Although many scales used in oral testing are of a global nature, there have been attempts to produce component rating scales such as those in the FSI oral interview procedure. All of these component rating scales contain similar concepts of what goes to make up the superordinate construct "oral ability," two of which are invariably the constructs "fluency" and "accuracy".

The constructs of fluency and accuracy have what may be termed an "intuitive reality" for most members of the teaching profession, who would claim that they could identify degrees of each in specific students. Indeed, much material development practice has grown up around "fluency practice" and "accuracy practice" and the concepts have been used in SLA (see pages 115 - 116). And yet, whether or not they are ratable in oral tests as valid components of oral ability has not been specifically investigated.

The development of a Fluency rating scale has been plagued by a lack of operational specificity since the very earliest FSI oral rating scales. In the FSI component scale for fluency, the scale constructors relied on vague concepts such as "slow and uneven speech" at band 2, and "hesitant and jerky speech" at band 3. At band 4 "groping for words" and "unevenness" is said to result in rephrasing, whilst by band 5 and 6 speech is
said to be "smooth".

The criterion of "hesitation" is also frequently found within rating scales of fluency. Some rating scales rely on the notion of hesitation alone, despite the fact that the phenomenon of hesitation is not well understood.

Two examples will suffice to highlight the point. These are taken from the University of Cambridge Local Examinations Syndicate First Certificate (FCE) and Certificate of Proficiency (CPE) examinations.

### The FCE Fluency rating scale

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Comfortable and natural speed and rhythm in everyday contexts, though there may be some hesitation when speaking in more abstract topics.</td>
</tr>
<tr>
<td>4</td>
<td>In everyday context speaks with minimal hesitation. Hesitation when discussing abstract topics, but does not demand unreasonable patience of the listener.</td>
</tr>
<tr>
<td>3</td>
<td>Does not hesitate unreasonably in everyday contexts, though may experience some difficulty with more abstract topics.</td>
</tr>
<tr>
<td>2</td>
<td>Unacceptable hesitation, even in everyday contexts.</td>
</tr>
<tr>
<td>1</td>
<td>Speech very disconnected.</td>
</tr>
<tr>
<td>0</td>
<td>Not capable of connected speech.</td>
</tr>
</tbody>
</table>

### The CPE Fluency rating scale

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Virtually native-speaker speed and rhythm, and coherent presentation of thoughts, in all contexts.</td>
</tr>
<tr>
<td>4</td>
<td>Foreign, but with minimal hesitation in all contexts.</td>
</tr>
<tr>
<td>3</td>
<td>Minimal hesitation in everyday contexts, but some hesitation when discussing more abstract topics, though not such as to demand unreasonable patience of the listener.</td>
</tr>
</tbody>
</table>
Hesitation not unreasonable in everyday contexts, but impedes understanding on more abstract topics.

Speaks haltingly even in everyday contexts.

Not capable of connected speech.

It would appear that the only theory (if theory it can be called) underlying these two scales amounts to the probably untested assumptions that (i) the better a student's English the less hesitation will be evident in performance, (ii) hesitation is much more likely when talking about "abstract" topics than "everyday" topics and, given the CPE band 5 descriptor, (iii) native speakers do not hesitate a great deal.

Fluent speech is often "disconnected", and "speed" and "rhythm" remain undefined: an analysis of native speaker talk reveals that unless the speech is deliberately pre-planned, hesitations and reformulations will abound (Fulcher, 1987). In the 16 brief conversational extracts in Crystal and Davy (1975) 146 examples of the hesitation marker/filler "er(m)" were found, showing that assumption (iii) above is clearly not the case.

If the third assumption is false, it follows that the first assumption is also likely to be false, but would benefit from empirical investigation. The real question is why the speech of both natives and foreign learners contains hesitation. The second assumption is merely baffling. Apart from the fact that "everyday" and
"abstract" are not defined but left to the intuition of the interviewer, there is no empirical evidence to suggest that talking about, say, politics in the Third World would prompt more hesitation than describing one's kitchen. The opposite may very well be true, as in courses leading to the FCE and CPE examinations students often discuss such subjects—subjects on which they might reasonably be expected to have some opinion. Not many native speakers or foreign learners find it easy to describe a ball-point pen (a "simple" topic in the FSI oral interview)!

Studies of hesitation phenomena have, however, suggested that time spent in hesitation is directly linked to "cognitive processing time" (Butterworth, 1975: 84), and that lexical selection may be one of the most important factors which influence hesitation. Chaudron (1988: 70) also suggests that pauses are often likely to be linked to planning phenomena. In a study of monitoring and repair in speech, Levelt (1983) found that 38% of all repairs were of a lexical nature, although he acknowledged that it was often extremely difficult to decide if the repaired lexical item was a real "error" or whether the speakers were making their utterance semantically more accurate (ibid., 54).

These studies suggest that the simplistic approach adopted by the developers of rating scales may result in the scales and band descriptors being invalid.
Section 2: A working definition of Fluency

Initially it is important to decide on what phenomena which can be observed within speech constitute an interruption in the fluency of speech. For the purposes of this study six phenomena were isolated as potentially interrupting fluency from an analysis of the transcripts of twenty one oral interviews, probably affecting the score given by an interviewer during an oral test. (It must be stressed that these phenomena are observational in that they occur on the surface of speech (verbal phenomena), and their description in no way constitutes an explanation of fluency. To provide explanations or interpretations of what is observed will be discussed shortly). The categories of phenomena generated in this study do not differ from the literature on fluency which discusses verbal aspects of performance which interrupt fluency (Grosjean, 1980; Hieke, 1985).

The data itself did not suggest any further phenomena which might have been investigated. The phenomena (or "exponents") are:

1. Fillers such as "er(m)".
2. The repetition of the first syllable of a word or a full word.
3. The negotiation of reference indicated by the re-selection of referring devices.
4. The re-selection of lexical items.
5. Anacolouthon.
6. Longer pauses, indicated in the transcripts and examples as two or three colons.

Category 6 needs special comment as this is an area related to the observation of fluency related phenomena which has recently seen much controversy. Studies involving the analysis of pauses in the second language literature have mainly concerned themselves with teacher speech and the role of pauses in adapting to communication with non-native speakers/learners (Ellis, 1985a: 145 - 146; Chaudron, 1988: 69 - 70), the suggestion being that speech rate is slower and the length of pauses greater in native speaker - non-native speaker communication similar to that observed between adults and children (Gais, 1977 as discussed in Allwright, 1988: 215), and that these phenomena help the comprehension of non-native speakers. Most of these studies do not use spectograph analysis, and the methodology used has come under severe criticism for lack of accuracy in measurement and drawing unsubstantiated conclusions from poor quality data (Griffiths, 1991).

Here, we are concerned with the effect that the impression of hesitation phenomena exhibited by non-native speakers under test conditions, including pauses, have upon raters. It is acknowledged that Griffiths' criticism of L2 research in the field of what he calls "pausology" is equally applicable to the way in which pauses in Category 6 of this data were observed and
recorded, that is, in terms of the impression created for raters. However, in defence of the analysis of this category, it must be said that the purpose of the analysis is somewhat different to that carried out in the currently available literature: we are not claiming that rate of speech and number of pauses is directly linked to comprehension, which after all is the major issue at stake. What is claimed is that the length and, perhaps, the nature of pauses in an oral test may be related to the impression given to the rater regarding the proficiency of the student and hence be related to the score which is awarded. For this purpose, precise measurement of pause length as advocated by Griffiths does not seem - at least for the present - to be a methodological consideration of primary importance.

With regard to all of the 6 categories of observational phenomena isolated as being related to the impression of fluency, it needs to be stressed that other factors not taken into account here, such as accent and propositional content, may go to provide a cumulative impression of the fluency of a speaker. The approach taken here has been to ignore phonological aspects of production which are most often treated separately in the testing literature and given a separate scale in component rating instruments. Other aspects of discourse such as the quality of propositional content are excluded because in an oral language test it is at least hoped that the student would be rated on language ability and
not the ability to develop an argument. Quality of propositional content would necessarily be linked to the topic of the task, thus favouring some students rather than others. It is therefore argued that a quality of propositional content rating scale would be more appropriate for an ESP oral test in which it might be expected that each student should be able to develop an appropriate argument in the L2, perhaps for attendance on a postgraduate course in that language. For a general test of oral proficiency including aspects of assessment which are likely to be task or topic specific could easily introduce further bias into the testing process. Further, it could very well be argued that the quality of propositional content does not represent an observed verbal phenomenon as described here. However, an attempt has been made to relate aspects of hesitation phenomena to attempts to develop propositional content in explanatory category 2 below, although the quality of the propositional content is seen to be outside the realms of an oral language test.

It is, however, not being claimed that the six categories of phenomenon isolated for study represent a complete and exhaustive definition of verbal elements which may be observed in the study of fluency. Indeed, it will be seen (pages 163 - 164) that other observable verbal elements were later considered to be worth investigation, and there may be others which have not been taken into account in this study.
Section 3: Data and analysis of categories

The data which was used for this study came from twenty one oral interviews conducted in Cyprus with Greek speaking learners of English. The students were selected at random from students taking the ELTS test at the British Council, Nicosia, in 1989. As it was considered important to collect the data under very similar circumstances, it was necessary that they all be interviewed by the same person under the same conditions, as far as possible, in the shortest amount of time possible. At the time, twenty one students was the maximum number that the interviewer could test within a reasonably small period of time. All interviews were transcribed by the researcher, and a coded transcript of one interview is provided in Appendix 4.12. The twenty one interviews were all conducted in the same room using the same interviewer, on three separate occasions within the space of one and a half weeks in 1989. This was done in order to control for test method facets such as physical environment and the style of the interviewer as much as possible. All the interviews were transcribed for later analysis in conjunction with the recorded tapes of the interviews.

Examples of phenomena which fell into the observational categories were isolated by the researcher from the transcripts and the tapes of the interviews, and then it was necessary to interpret the examples in terms of the explanatory categories. Where it was possible to
specify linguistic criteria for identifying one of the phenomena concordancing facilities were used to extract all occurrences from the transcripts in order to isolate consistencies in the data.

The verbal phenomena or exponents of fluency as described in the categories above were isolated and each example was placed into one of eight explanatory categories. The exponents are observable verbal elements of speech which are said to constitute an interruption in fluency. The explanatory categories represent an attempt to say why fluency appears to be interrupted (or not, as the case may be) rather than merely recording tallies of say, examples of pauses of different lengths. It is necessary to introduce explanatory categories in addition to the categories of verbal phenomena because the same phenomena may be interpreted by raters in different ways. For example, some longer pauses may be interpreted by raters as indicating that communication has broken down, and the student would therefore be penalised on the rating scale. However, it is quite possible that the rater will consider certain pauses as being "thinking time" in which the student is seriously considering the content of the next utterance, and thus be prepared to award a higher band for natural language behaviour (Meredith, 1978). The explanatory categories were generated by the researcher from the data itself, rather than attempting to produce explanatory categories on an a-priori basis.
Unless it is possible to claim a one-to-one relationship between the categories of phenomena and the explanatory categories into which they are placed, it is always possible for the researcher to be challenged regarding the interpretation of the data. It is therefore important to look at problems which arise in a context where observed phenomena described as interrupting fluency are then transferred to explanatory categories for the use in actual scale development.

Long (1983: 3) makes clear the difficulty faced by an approach to scale development which depends upon the analysis of actual learner speech when he comments on similar difficulties faced by any researcher who is working with spoken data. Although researchers recognise "the bias in one person reporting events" they are prepared to do this in order to attempt a meaningful description of the phenomena which they observe. In the context of this study, it would be possible merely to observe the six phenomena which have been listed above, and to use the number of occurrences as a guide to scale construction. Such a scale would, it is suggested, be termed a "low inference" scale, which relates to specific observed behaviour during an oral test. It may be hypothesized that this would lead to increased reliability in using the scale developed, but would have the disadvantage of making it trivial in the sense that no inferences are actually being made about why the observed behaviour is occurring. When discussing the use
of low-inference observational categories in connection to classroom observation, Long comments:

"The increase in reliability that these are designed to bring is largely mythical...and comes at a high price. A focus on overt behaviour may, for example, "reliably" preempt the explanatory power of a study by precluding consideration of participants' intentions." (Long, 1983: 13)

Long was, of course, attempting to deal with data derived from classroom observations, whilst this study is attempting to deal with the development of valid rating scales in oral tests. However, the example of the difference between low and high inference interpretation which Long provides may help to clarify the issue at stake. Long (1983: 13 - 14) provides the following extract from an intermediate class:

1. T: ...OK? Chemical pollution. OK.
2. S4: (Yawning) O o o.
3. T: Trousers! Alright, Carlos [S4], do you wear trousers?
4. S4: Always. All my life.
5. SS: (laughter)
6. T: Always. You've worn, I have...
7. S4: Eh wear wear (inaudible).
8. T: I have... well, do you wear trousers?
10. SS: I wear, I wear.
12. T: Yes, you do. What's how do you say that word?
15. S4: Trousers.
17. S4: Trousers.
19. T: Mm hm. Have you got trousers on?
20. S3: Yes, I have.
21. T: What kind?
22. S3: Jeans.
23. T: Jeans...Say the word jeans. Jeans.
25. T: Jeans.
27. T: Jeans.
28. S1: Jeans.

Long argues that current forms of analysis of classroom interaction which depend on low inference categories could not capture the observation made by Long himself that the exchanges in this extract were motivated by S4 who overtly expressed boredom, and by the use of the joke in line 4 was directly challenging the "usefulness of the lesson" (Long, 1983: 14). The drilling which follows is used by the teacher to reassert authority and not to drill new vocabulary. This interpretation of the data is "high inference" in that explanation is introduced which is not strictly data dependent. It may even be said that a theory is being introduced to account for the data as observed. However, it must be remembered that a transcript is a reduction of the entire data, and with the additional use of the recording of the lesson, the high level inferences would seem to be much more legitimate than through mere analysis of the transcript.

In the explanatory categories which have been generated for the purpose of the Fluency rating scale developed here, there is no one-to-one relationship between the observed phenomena listed in the six
categories above and the explanatory categories, as an attempt has been made to interpret the purpose of the interruption to fluency in each case. As such, it tends to be a high inference scale which, while being more open to criticism and certainly more difficult to create, does offer the possibility of non-trivial description of student performance.

Section 4: Eight Explanatory Categories

The data suggested that observed interruptions in fluency could be accounted for by eight categories. These are:

1. End of turn pauses: pauses indicating the end of a turn.
2. Content planning hesitation: pauses which appear to allow the student to plan the content of the next utterance.
3. Grammatical planning hesitation: pauses which appear to allow the student to plan the form of the next utterance.
4. Addition of examples, counter-examples, or reasons to support a point of view: these pauses are used as an oral parenthesis before adding extra information to an argument or point of view, or break up a list of examples.
5. Expressing lexical uncertainty: pauses which mark searching for a word or expression which has been temporarily forgotten or about which the student is uncertain.
6. Grammatical and/or lexical repair: hesitation phenomena which appear to be associated with self-correction.
7. Expressing propositional uncertainty: hesitation phenomena which appear to mark uncertainty in the views which are being expressed.
8. Misunderstanding or breakdown in communication.
Although it has been stated that the above explanatory categories are, generally, high inference categories, they do vary in the amount of inference which is necessary on the part of the researcher. Category one, for example, is fairly low inference in that it merely involves observing those pauses which occur at the end of turns where propositional content is coherent. Category two, on the other hand, is high inference in that one must attempt to decide when hesitation occurs because the learner is deciding what to say next, as opposed to how to say it, which is the definition of Category three. This means that when an example of one of the six phenomena described as constituting aspects of fluency is observed, it must be decided into which of the eight explanatory categories to place it.

Any non-trivial descriptive system will suffer from such problems of indeterminacy in the developmental stage. (When a set of observations or interpretations are said to suffer from "indeterminacy" we mean that there is doubt about the degree to which they represent a "fact of the matter" state in the world which is independent of the research which generates them (Miller and Fredericks, 1991: 4 - 5)). However, when developing a rating scale the adequacy of the initial analysis and construction of band descriptors based on that analysis comes in the trialling of the scale. That is, if the initial analysis and the band descriptors which can be constructed on that basis are valid descriptions of levels of fluency, then
validity coefficients for those bands will be acceptably high thus providing evidence that a data-driven approach to rating scale development was useful. On the other hand, if the band descriptors do not relate to observable behaviour and if raters cannot apply the concepts in the practice of rating students, then the validity of the scale would be low, and suggest that a data-driven approach to scale development may need to be trivial and thus possibly not preferable to a-priori scale development.

The eight explanatory categories used are described here in full, with examples.

**Category One: End of Turn Pauses**

The most frequent use of extended pauses in the data was at the end of a student's turn (in the examples, a "B" turn). Example 4.1 is taken from a student awarded a band 4 of the ELTS rating scale.

**Example 4.1.**

\[ B>...fifteen thousand womans went to er: prisons :: er :::: A> and since then B> yeah A> has there been an increase or a decrease..." \]

In these cases, it would appear that the students pause because they are not able (or willing) to continue speaking. The interviewer does not begin his turn immediately as he is waiting for the student to continue. The interviewer in the oral test appears to be highly
sensitive to the possibility that the student needs time to plan what is going to be said next, and therefore the amount of overlapping speech may be much less than in less formal interaction, and the amount of silence between turns increases, something which would be highly embarrassing in informal talk (Sacks, Schegloff and Jefferson, 1974).

When turning to the high scoring students (arbitrarily defined as band 7 and upwards on the ELTS rating scale), we discover that the amount of this type of hesitation is somewhat different in format. For example:

Example 4.2.
A> ...it sounds absolutely [fascinating B> it is]:: A> actually...

In this example the B turn overlaps with the previous A turn, and the pause allows B to continue the conversation if B so desires. It is not taken up, and so A continues not by asking another question to get the student to complete the proposition (as it is already complete!), but by continuing the conversation. In oral tests, the experienced interviewer does not really wish to keep a turn for any length of time. When the conversation is more "natural" with higher ability students, the interviewer appears to take more responsibility for the existence of potentially embarrassing pauses, as is shown in example 4.3.
Example 4.3.
B> ...but that takes a while::: A> so: my next question is...

It would seem that in example 4.3, A is embarrassed at having allowed such a long pause at the end of the student’s turn, and therefore uses a discourse marker of organization at the beginning of the turn, and an overt signal that he is in a dominant social position by stating that he is there to ask questions. This particular example may be compared with phenomena such as those discussed by Long (1983: 14) in which teachers (and, we assume interviewers) negotiate discourse in such a way as to maintain their social role in the interaction.

Category Two: Content Planning Hesitation

Hesitation does seem to have a significant psycholinguistic role in that it allows the speaker time to formulate a plan for the next chunk of language. One aspect of this planning is the content of the utterance.

Example 4.4.
B> ...some in the lungs:: from the lungs it goes er through the reticuloendothelial system...

Example 4.5.
A> ...in this way B> what reason:: I I should say it must be the er: our er: con er contribution to...

In example 4.4 the student is considering the next
step in a process which is being planned with the help of an input diagram. In example 4.5, the student overtly indicates that content planning is going on, as the question which the interviewer has asked is summarised before the pause and the beginning of the answer. This seems to be a case of the student "thinking out loud" or introspecting, which enables the researcher to classify the example fairly easily.

In example 4.6, the student appears to be considering her future plans and the time used after the filler is to plan the content of the next part of the message.

Example 4.6.
B> I think that we'll we will be together A hm hm B> and er:: after that I want to: to be a judge...

Category Three: Grammatical Forward Planning

Grammatical forward planning occurs when the student may know what to say, but does not know how to say it. Thus, the planning stage needs time, and the execution may also contain other hesitations and structural reformulations.

Example 4.7.
A> ...kind of a job do you want:: B> when I chose this subject: I:: I chose this sub I chose this subject because I think I can...

In this example, it appears that the student wishes
to express the view that the subject of study was chosen because of the ease with which (she goes on to say) it would be possible to get any kind of job. The purpose clause, beginning with "because..." is therefore the most important part of the message. However, having begun with "when" the students seems to find difficulty in actually formulating the message intended as it requires a purpose clause which cannot be used given the colligational restrictions which have been activated, unless the student can use the cohesive form "I did so (because)". Hence, the medium length pause after "I" and the two reformulations. In the process the use of "when" is avoided, and the purpose clause is successfully introduced.

Example 4.8.
B> ...er in the fields er: the house:: they they have er they have their work they gets like er: like er:...

In example 4.8 in the formulation of the utterance after "house::" the student gets across the message because of the collocation of "work" with "house" and "fields", whilst the hesitation associated with the multiple use of "they" (referring to girls who live in villages) indicates that the appropriate grammatical structures to produce an accurate utterance may not be available even though the student is searching for them.

In example 4.9 below the repetition of items at the beginning of the utterance indicates that planning is
Example 4.9.
B> I'll try to get back to Cyprus to make some more lectures to others.

Category Four: Addition of examples, counter examples, or reasons to support a point of view

It was observed that a pause or a filled pause often precedes the addition of an example or reason when the student is presenting an argument, adding to or supporting what has already been said, as in example 4.10.

Example 4.10.
B> ...the television crimes and they are trying to mime them: and er: the newspapers as well they don't erm: stick to the facts...

It may also be noted that in the clause after the filled pause here the phrase "as well" overtly signals the additional nature of the information to support the argument. Students of higher ability were noticed to be particularly adept at using pauses or fillers to introduce an additional element to their argument; in the case of example 4.10, copying crimes on television is one cause of the increasing crime rate, but another cause may be the sensationalism of reporting in the written media. Another example of the simple addition of examples is given in example 4.11.
Example 4.11.
B> ...diarrhoea vomiting A> hm B> headaches and er: ataxia erm:: convulsions sweating A> it sounds pretty horrible B> yeah it is...

An example of a counter argument is given in example 4.12.

Example 4.12.
B> ...I was interesting about this subject:: of course erm: you might erm: think that er since I: I am already...

The student in example 4.12 provides a reason why he wishes to pursue a subject at university, and then goes on to give a reason why it may be thought that he would not be interested (namely that he is already employed in the area) and then goes on to justify his original interest.

Amongst the students who received lower band scores on the ELTS rating scale very few occurrences of giving examples, counter examples or reasons were observed. When they did occur the most common use was that of adding to an utterance in order to give content to a general word which otherwise would have remained empty. Such general words are often referred to as "delexical" in that they do not of themselves contain any specific meaning, and have to be filled out with reference to context or other parts of the discourse (Winter, 1978). This is the case in example 4.13, where the delexical "routine" is filled out with the specific information of going to the office.

143
and (after this) doing the work which is assigned by a superior.

Example 4.13.
B> it's the same thing every day the same routine:: er I've got to go every morning to the office...

**Category Five: Expressing Lexical Uncertainty**

When choice of a lexical item was a problem for a student in the higher bands of the ELTS rating scale there was a tendency for it to be overtly marked, whilst not causing any strain on the interaction. Examples 4.14 and 4.15 demonstrate this.

Example 4.14.
B> ...rate is increasing A> hm hm B> of girls er committing:: er loneliness or what what do you call it: I mean: committing crimes...

Example 4.15.
B> ...physical disaster than er there is no erm: civil liability:: I think that is how A> hm hm B> it is called: but er in in case er...

In example 4.14 the lexical item needed does not come to mind, but the student still continues and explains what he wishes to say in another way. This use of circumlocution marks the student as a competent communicator. In example 4.15 the student appears to need time to consider the use of the expression "civil liability" and overtly marks concern over the selection, but continues once the interviewer has indicated that
this makes sense and that the student has communicated his intention.

Example 4.16 is from a student in the lower bands, and is an example of the abandonment of the word when it is realised that the correct form is not known for the context.

Example 4.16.
B> one er English legal system philosophy: erm: govern er:: and er: er tort contract A> hm B> and it’s all...

At times, the failure to retrieve a lexical item required and the inability of the student to circumlocute may lead to difficulty in retrieving a message, as in example 4.17 where the message is unclear.

Example 4.17.
B> ...they they have a choice with going to er: with erm:: er stranger ma with people A> hm B> going to drugs: to: with with er...

Category Six: Grammatical or Lexical Repair

When analyzing the data the separation of grammatical and lexical repair was found to be virtually impossible. The two appear to overlap to such an extent that to separate individual examples would have involved what was considered to be an undue amount of subjective decision. Grammatical and lexical repair were thus treated as a single category.

In example 4.18 the student uses an incorrect
preposition, but repairs this in a repetition of the entire prepositional phrase, and appears to consider the next utterance carefully, repeating the correct preposition twice. Example 4.19 is interpreted similarly.

Example 4.18.
B> ...is stored in in the body er on the bones: in the bones: and::: if the:: plutonium is stored in the body in in the bones A> hm hm B> ...

Example 4.19.
B> ...erm: in the other on the other hand::: erm I have my clearly medical duties A> hm hm B> in diagnosis...

In example 4.20 we have an example of anacoluthon with a slight pause. After two false starts in this case the student realises a structure which will allow the message to be communicated. The two previous starts would probably have led to trouble in completing the proposition.

Example 4.20.
B> ...but I think in Cyprus: er in we are not very: the girls are not er crim criminals

At times the accuracy of the utterance breaks down completely, and a pause probably indicates that re-planning is taking place, as in example 4.21.

Example 4.21.
B> ...essay: er about er: the metals: er but it er: I couldn’t er:: the subject was about er: which metals can be used er: in a car...
It is suggested that in example 4.21 the repair is fairly successful: following the phenomenon of anacolouthon an appropriate grammatical structure to complete the sentence is found.

**Category Seven: Expressing Propositional Uncertainty.**

It was observed that students tended to hesitate when expressing uncertainty about a point of view or an argument which was being presented. This may be connected to the fact that the interviewer is the socially dominant partner in the interaction, as observed by Channell (1985). The student may attempt to avoid strong claims where points of view are presented as fact and potentially seen as challenging a dominant partner; this would not occur when the interviewer is offered the possibility of disagreeing.

Examples 4.22 - 4.24 falling into this category were clearly marked not only with hesitation phenomena but also with one of the following items which makes categorisation somewhat easier: perhaps, maybe, think, believe, don’t know, or sometimes.

**Example 4.22.**
B> ...so we have no time to: study: for our school: and erm:: er but I think that er English schools that er all: are the best...

**Example 4.23.**
A> ...what kind of a job would you be looking for B> erm:: maybe in er in the government er the best [place A> hm] B> to work...
Example 4.24.
B> ...communication between students and: er: teachers A> yeah B> erm:: well I think that er (inaudible) is er good A> hm...

The co-occurrence of the hesitation phenomena with the overt lexical items isolated gives to the utterance a sense of reservation on the part of the speaker which would admit the possibility of a challenge from the interlocutor.

**Category 8: Misunderstanding or Breakdown in Communication.**

Examples from the data which were placed into this category came entirely from those less able students who received a band score of 4 or 5 on the ELTS rating scale.

Example 4.25.
A> ...has there been an increase or a decrease in crime::: B> increase:: in crimes:: A> has there been an increase or a decrease in the crime rate...

In example 4.25 the student has not understood the question, and echoes part of the interviewer’s utterance not as an answer but as an indication that the question has not been understood. Without the hesitation it might have been taken as an answer. In example 4.26 the student explicitly acknowledges that he has not understood the question.
Section 5: **Discriminant analysis**

In order to create a Fluency rating scale in which the band descriptors are generated by the data, it is necessary to make tallies of the observations classified into each of the eight explanatory categories described above, and compute discriminant functions for the eight categories to see whether or not, taken together, they are able to predict membership of the ELTS band scores which the learners who provided the data actually received in the oral test. Discriminant analysis allows the researcher to investigate (a) the extent to which all categories together, and each individual category taken separately, discriminate between the students, and (b) the extent to which students would have been reliably placed in the bands which they actually received in the ELTS oral interview had they been rated on the categories developed (Crocker and Algina, 1986: 256 - 263).

Clearly, an assumption is being made in the use of such a technique in the development of a scale: that the students were at least rank ordered appropriately in terms of ability by the ELTS oral interview. For the moment, it will be assumed that this was indeed the case, although comparison between the ELTS rating scale and the rating scales developed in this study will be made in
In the discriminant analysis tallies of the phenomena described within each of the eight explanatory categories were regressed on the actual band scores of the students, which operated as the categorial variable in the analysis. It was assumed that if it proved possible to discriminate accurately between students using the categories isolated, it should also prove possible to develop a rating scale on the basis of the data which could then be validated in use on other groups of students. That is, if the categories are found to discriminate between students placed in certain bands they may then be used to create verbal descriptors for new bands by returning to the description which generated the categories. The new descriptors thus contain definitions which are directly linked to actual L₂ production in test conditions.

Only one student was awarded a band 8 and one a band 9, and so these were pooled during the analysis. The frequency distribution of scores on the ELTS test are contained in Table 4.1 in the Appendix to Chapter 4.

The multivariate results presented in Table 4.2 in the Appendix to Chapter 4 were produced to predict, from all observations on all categories, the students' actual oral test scores. This shows the degree to which all categories taken together are capable of discriminating between the students. Wilks' Lambda is the test of the multivariate hypothesis that subjects can be divided into
groups from observations (Wilkinson, 1988: 538). The result of the multivariate analysis is significant at $p = .02$, indicating that when taken together the categories do seem to discriminate well between students. It will also be seen that the first discriminant function in the analysis is significant, but the second and subsequent functions are not.

Table 4.3 provides simple correlations between each of the categories and the ELTS band for the 21 students. Although these correlations show a clear reduction in hesitation for grammatical planning and misunderstanding or communicative breakdown (categories 3 and 8) and an increase in the use of hesitation to introduce examples or arguments (category 4) with increasing student ability, these correlations mask other differences between ability levels. These are discussed with reference to the means for each band on each category on pages 153 to 159 below, where it will be demonstrated that we are not in fact dealing with linear phenomena.

Finally, by using discriminant analysis it is possible to analyze the relationship between the band score actually awarded to each of the students and the band score which would be predicted on the basis of the categories which are thought to represent aspects of perceptible interruptions in the fluency of speech. The results of such prediction is presented in Figure 4.1.
In Figure 4.1, band 1 represents a band 4 on the ELTS rating scale, band 2 an ELTS band 5, and so on, but with band 5 being the conflated bands 8 and 9. In the left hand column we have the bands which were actually awarded in the ELTS oral test, and these are compared with the bands which would have been awarded if they had been awarded on the basis of those explanatory categories which proved significant discriminators in the discriminant analysis (Wilkinson, 1988: 589). Thus, for example, eight students were actually awarded a band 3 (ELTS band 6), whereas if these eight students had been awarded their score on the basis of the significantly discriminating explanatory categories only 7 would have been awarded a band 3 and one would have been awarded a band 4.

It may be seen from Figure 4.1 that only one candidate would have been given a different band score.

<table>
<thead>
<tr>
<th>Band Awarded</th>
<th>Predicted Band</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
As prediction appears from this data to be very accurate, this may be taken as an indication that the data-driven approach to scale development is indeed worth pursing.

Section 6: Producing Fluency rating scale descriptors

Having discovered from the discriminant analysis that a single discriminant function allows prediction of the ELTS band score from the eight categories described in Section 4 above, we may turn to an analysis of the pattern of means on those categories by band score. The simple correlation coefficients given in Table 4.3 show that there is a correlation between category and band score and category for categories 3, 4 and 8. However, it was argued on page 151 that the other correlation coefficients mask very real patterns in the data which can be used in the construction of band descriptors for a Fluency rating scale. The mean scores of students in each band for each category are given in Tables 4.4 to 4.11 in the Appendix to Chapter 4, and Figures 4.2 to 4.10 on pages 154 - 157 are the graphs of these means by band.

Figures 4.4, 4.5 and 4.10 clearly show the linear relationship between category and band score. In each other case there is a non-linear relationship which nevertheless serves to discriminate between bands on the rating scale.

The process of producing scale descriptors can now be attempted by returning to the description of the categories which can be seen to discriminate between
Fig. 4.2. Category 1 means

![Graph showing means for Group 1 to Group 5.]

Fig. 4.3. Category 2 means

![Graph showing means for Group 1 to Group 5.]
Fig 4.4. Category 3 means

Fig 4.5. Category 4 means
Fig. 4.6. Category 5 means

Fig. 4.7. Category 6 means
Fig. 4.8. Category 7 means

Fig. 4.9. Category 8 means
bands and translating the information for use in new bands which would discriminate between these and other students.

In order to exemplify the process we will examine Category 1, end of turn pauses, in some detail. On page 138 it was observed from the analysis of the transcript data and tapes that end of turn pauses were most frequent in students of lower ability and higher ability. Figure 4.2 shows this "U" shape most clearly. In the case of lower ability students the pause was interpreted by the researcher as an indication of an inability to continue speaking from the evidence provided by the transcripts, the audio tapes and recollections of the interview itself. In the case of higher ability students, the pauses indicated an end to a propositionally complete utterance and an end of the student turn, to which the interviewer did not respond immediately because the next question or prompt had not been prepared. End of turn pauses indicate the end of a turn in both cases, but are interpreted as occurring for different reasons.

This category can therefore be used to distinguish band 1 and band 5 students from those in bands 2, 3, and 4, but in a way which also allows a distinction to be drawn between band 1 and band 5. The rho of -0.023 for the relationship between band and category 1 in Table 4.3 hides this important distinction.

Band 1 in the Fluency rating scale therefore states that students often "dry up" or cannot continue a
conversation, whilst the Band 5 descriptor states that students often respond to prompts or questions so quickly that the interviewer appears not to be prepared for the next prompt or question, and that there is often a pause as a result.

Such non-linear interpretations are necessary for all categories which have low correlation coefficients in Table 4.3. It must therefore be stressed that in the construction of the band descriptors for the Fluency rating scale it was necessary to interpret the explanatory categories in the light of the data.

The Fluency rating scale contained five bands which were generated by the data, labelled bands 1 to 5. Two further bands were attached to the scale: band 0 was merely described as "less fluent" than band 1, whilst band 6 was described as "more fluent" than band 5. The purpose of this was twofold. Firstly, it is not being claimed that the rating scale taps the entire range of potential fluency. That is, bands 1 and bands 5 are not seen as absolute extremes. It was merely not possible to describe anything less or more fluent from the data, and this avoids the further problem of suggesting that there is such a thing as zero and perfect proficiency in fluency. Secondly, it was hypothesized that raters would tend to avoid the highest and lowest bands on the scale, and thus the inclusion of two additional bands would mean that there would be a greater chance of raters using the full range of bands which can be described.
The Fluency rating scale is given in full below, and an explanation of the relationship of the descriptors to the eight explanatory categories is contained in [square brackets]. It will also be noted that some parts of the descriptors are contained in (round brackets), as these do not appear to be related in anyway whatsoever to the preceding discussion. This will be taken up later in Section 7 below.

The Fluency Rating Scale

Band 0

Candidates in band 0 do not reach the required standard to be placed in band 1.

Band 1

The candidate frequently pauses in speech before completing the propositional intention of the utterance, causing the interviewer to ask additional questions and/or make comments in order to continue the conversation [categories 1 and 8]. (Utterances tend to be short), and there is little evidence of candidates taking time to plan the content of the utterance in advance of speaking [category 2]. However, hesitation is frequently evident when the candidate has to plan the utterance grammatically [category 3]. This often involves the repetition of items, long pauses, and the reformulation of sentences.

Misunderstanding of the interviewer’s questions or comments is fairly frequent, and the candidate sometimes cannot respond at all, or dries up part way through the answer [categories 1 and 8]. (Single word responses followed by pauses are common), forcing the interviewer to encourage further contribution. It is rare for a band 1 candidate to be able to give examples, counter examples or reasons, to support a view expressed [category 4].

Pausing for grammatical and lexical repair is evident i.e., selection of a new word or structure when it is realised that an utterance is not accurate or cannot be completed accurately [category 6].

Candidates at band 1 may pause because of difficulty in
retrieving a word, but when this happens will usually abandon the message rather than attempt to circumlocute. It is rare for a band 1 candidate to express uncertainty regarding choice of lexis or the propositional content of the message [category 5]. (The message itself is often simple).

Band 2

A band 2 candidate will almost always be able to complete the propositional intention of an utterance once started, causing no strain on the interviewer by expecting him/her to maintain the interaction [category 8]. However, just like a band 1 candidate, a band 2 candidate will frequently misunderstand the interviewer's question or be completely unable to respond to the interviewer's question, requiring the interviewer to repeat the question or clarify what he/she wishes the candidate to do [category 8]. Similarly, (single word responses are common), forcing the interviewer to encourage further contribution.

Although the candidate will spend less time pausing to plan the grammar of an utterance, it will be observed that there are many occasions on which the candidate will reformulate an utterance having begun using one grammatical pattern and conclude with a different form [categories 3 and 6]. Similarly, with lexis, there will be evidence that the candidate pauses to search for an appropriate lexical item and, if it is not available, will make some attempt to circumlocute even if this is not very successful [categories 5 and 6]. From time to time a band 2 candidate may pause to consider giving an example, counter example, or reason for a point of view. However, this will be infrequent and when it does occur the example or reason may be expressed in very simplistic terms and may lack relevance to the topic [category 4].

Band 3

A candidate in band 3 will hardly ever misunderstand a question or be unable to respond to a question from the interviewer. On the odd occasion when it does happen a band 3 candidate will almost always ask for clarification from the interviewer [category 8].

Most pauses in the speech of a band 3 candidate will occur when they require "thinking time" in order to provide a propositionally appropriate utterance [category 2]. Time is sometimes needed to plan a sentence grammatically in advance, especially after making an error which the candidate then rephrases [category 3].

A band 3 candidate is very conscious of his/her use of lexis, and often pauses to think about the word which has
been used, or to select another which they consider to be better in the context. The candidate may even question the interviewer overtly regarding the appropriacy of the word which has been chosen [category 5].

Often candidates in this band will give examples, counter examples or reasons to support their point of view [category 4].

(At band 3 and above there is an increasing tendency for candidates to use "backchanneling" - the use of "hm" or "yeah" - when the interviewer is talking, giving the interview a greater sense of normal conversation, although many better candidates still do not use this device).

**Band 4**

A band 4 candidate will only very rarely misunderstand a question of the interviewer, fail to respond, or dry up in the middle of an utterance [categories 1 and 8].

A candidate in this band will exhibit a much greater tendency than candidates in any other band to express doubt about what they are saying. They will often use words such as "maybe" and "perhaps" when presenting their own point of view or opinion [category 7]. More often than not, they will back up their opinion with examples or provide reasons for holding a certain belief [category 4]. They will pause frequently to consider exactly how to express the content of what they wish to say and how they will present their views [category 2]. (They will only rarely respond with a single word unless asked a polar question by the interviewer).

There will be far fewer pauses to consider the grammatical structure of an utterance [category 3] and pausing to consider the appropriacy of a lexical item chosen is rare [category 5]. A candidate in this band will reformulate a sentence from time to time if it is considered to be inaccurate or the grammar does not allow the candidate to complete the proposition which he/she wishes to express [category 6].

**Band 5**

A candidate at band 5 almost never misunderstands the interviewer, fails to respond, or dries up when speaking [categories 1 and 8]. The majority of pauses or hesitations which occur will be when the candidate is considering how to express a point of view or opinion [category 2], or how to support a point of view or opinion by providing appropriate examples or reasons [category 4]. However, a candidate at band 5 will not express uncertainty regarding these views or opinions as
frequently as a candidate at band 4, and so there are fewer hesitations when introducing new propositions [category 7].

Very rarely does a band 5 candidate have to pause to consider the grammatical structure of an utterance [category 3] and almost never hesitates regarding choice of lexis [category 5]. Band 5 candidates demonstrate a confidence in their ability to get things right the first time. Whilst they do sometimes pause to reformulate sentences this is always because they cannot put across the propositional content of their utterance without changing grammatical form [category 6].

It may be noticed by the interviewer that the candidate responds to questions and prompts so quickly and efficiently that the next question or prompt has not been prepared, resulting in a pause in the interview while the interviewer plans his/her next utterance [category 1].

Band 6

Candidates in band 6 reach a standard higher than that described in band 5.

Section 7: Discussion

It will be noted that the sections in descriptors which are contained in round brackets refer to the number of single word utterances and the use of backchanneling. Neither of these phenomena were included in the original definition of fluency, nor were they analyzed in the data when the Fluency rating scale was first drafted. Whether or not they properly belong within a Fluency rating scale may be contentious. However, when developing the Accuracy scale it was discovered that lower ability students tended not to make many "global errors" because the strings of speech which they produced were relatively short and the messages produced correspondingly simple, whilst single word utterances of higher ability students
included backchannels. Although the analysis of single word utterances and backchannels was carried out in relation to the Accuracy scale in order to look at the degree to which error rates related to amount of speech produced, it was felt that these phenomena were not strictly related to the accuracy of the student's speech, but to the perception of fluency in maintaining the flow of the conversation.

Single word utterances and backchanneling will not be discussed further here, as the data is presented and discussed on pages 199 - 200 in relation to the development of the Accuracy scale. It is however important to note that these parts of the descriptors were included during a revision process after both scales had initially been drafted.

The Fluency rating scale which has been developed here may well avoid two problems which have already been discussed. The first of these is the problem of falling into the trap of failing to define abilities in enough detail to make it impossible to investigate the validity of the scale. Secondly, the descriptors are specific enough to be relatable to actual language performance if we assume, for it must at the moment be an assumption, that the approach to the analysis of the original data used in this study to develop eight explanatory categories is a useful one.

Above all, should the scale prove to be reliable and valid, then it would be possible to make non-trivial
claims concerning the generalizability of the definitions in the band descriptors. One would also hope that it would be safe to generalize beyond students from the population from which the samples used in this research were drawn. This would appear, at the end of the day, to be the main purpose in attempting to generate descriptors from actual language data.

The data-driven approach to scale development is therefore different from traditional approaches in the FSI and ILR moulds, and also different from the "absolute" scales which researchers are currently investigating (Bachman and Clark, 1988; Bachman, 1990). The degree to which such a rating scale is successful can only be evaluated through trialling and validation studies, but initial evidence from discriminant analysis suggests that this approach to scale development is promising.
Chapter Five

The Development of an Accuracy Rating Scale

Section 1: Introduction

Research into rating accuracy of language in free oral production is a subject which would appear, with one or two exceptions, to have been neglected in the testing literature to date. Weir (1988: 89), for example, notes that in tests of production it is important to have detailed marking schemes to guide examiners (in public examinations) in rating candidates for the accuracy of language produced. However, he does not expand on this. Madsen (1983, 166 - 174) discusses differences between holistic and discrete scoring techniques for accuracy, maintaining that holistic scoring is preferable for experienced teachers for whom oral evaluation does not present any problems. To demonstrate an accuracy scale which he believes to be a good example of its type he uses that from the Test of English as a Second Language of the American Language Institute/Georgetown University (ALI/GU), which is general in terms of its descriptors:

5 points: Uses English with few (if any) noticeable errors of grammar or word order.

4 points: In general uses "good English," but with occasional grammatical or word-order errors which do not, however, obscure meaning (eg., "I am needing more English").

3 points: Meaning occasionally obscured by grammatical and/or word-order errors.
2 points: Grammatical usage and word-order definitely unsatisfactory: frequently needs to rephrase constructions and/or restricts himself to basic structural patterns (e.g., uses the simple present tense where he should use past or future).

1 point: Errors of grammar and word order make comprehension quite difficult.

0 points: Speech so full of grammatical and word-order errors as to be virtually unintelligible to "the man in the street."

Underhill (1987) goes somewhat further in expressing the view that marking oral performance is really quite simple. In his discussion of scoring (Underhill, 1987: 95 - 103) he does recommend the use of rating scales, but does not deal with accuracy other than to say that it should be assessed.

Baker (1989: 84 - 92) is somewhat more sensitive to the problems of rating accuracy. With reference to the ELTS and the FSI rating scales he notes that the descriptors (especially in the latter test) do not relate directly to the language the candidate is to produce. In his scale construction recommendations he states that developers should "make the scales themselves explicit and unambiguous and refer to readily observable aspects of the performance" (Baker, 1989: 87). Hughes (1989: 102) quotes from the Royal Society of Arts (RSA) intermediate oral interaction scales: "Grammatical/lexical accuracy is generally high, though some errors which do not destroy communication are acceptable." He notes, quite correctly, that:
"The reader may well feel that these descriptions of levels of ability are not sufficiently precise to use with any consistency. In fact, even where specification is more precise than this, the operationalisation of criterial levels has to be carried out in conjunction with samples of candidates’ performances."

It is precisely this point that has been taken seriously in the development of the accuracy scale to be described later.

These discussions of accuracy in the literature have gone little beyond the observation of Lado (1961: 246) that

"...rating scales of sample responses at each rating level are of course an improvement, but unless the rating scale tells the examiner exactly what elements to consider right and what to consider wrong, he is still left to his own opinions and listening habits at the moment of scoring."

Lado argued for explicit descriptors in just the same way that Baker and Hughes are currently doing. Matthews (1990) is essentially correct when she concludes that a major step forward in the testing of oral performance can only come about once attention is given to the criteria against which the performance is to be matched.

Some work has, however, been done in this area. Adams et. al. (1987) and Griffin et. al. (1988) report the development of a grammatical accuracy scale using
Rasch Partial Credit models in Australia based on a model suggested by Higgs and Clifford (1982) and the so-called Functional Trisection described by Liskin - Gasparro (1984c). Griffin et. al. (1988: 7) describe the six areas upon which their tests were developed. Firstly, a general objective was described for each test "item", and secondly a possible language element, namely, the most likely grammatical structure which would realise the general objective. Thirdly, the questions used in the elicitation techniques were described by question type, formula for construction and an example of a likely question. Fourthly, restrictions in formation in terms of vocabulary and context were included. Fifthly they provide instructions for obligatory procedures to be used by the examiners and finally scoring procedures. An example, from Griffin et. al. (1988: 10) is as follows:

General Objective: Students will be able to describe actions being performed at the present time.

Possible Language element: Present Continuous forms.

Question type: Restricted question.

Formula:  

```
<table>
<thead>
<tr>
<th>What</th>
<th>are</th>
<th>the</th>
<th>pronoun</th>
<th>&lt;doing&gt;</th>
<th>happening?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>is</td>
<td>these</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

where angular brackets = element to be replaced by a word or phrase, elements between straight lines are alternatives, and elements in square brackets are optional.

Example: What are these people doing? (a set of sketches are provided for the test item)
Restrictions: The verbs chosen should be selected on the basis of common actions familiar to the client.

Instructions:
(a) Place the stimulus in front of the client or indicate the room or other context to be identified.
(b) Elicit three (3) or more responses from the client.
(c) If the student falters after the instructions have been given, indicate specific actions to be identified.

Response Criteria:
0 - No continuous form used or understandable
1 - A number of continuous forms are used but production of the list involves hesitation and self correction. There is little or no fluency in the production of the list.
2 - A clear list of verbs in the continuous form.

Adams et. al. (1987: 21) report that by using Rasch fit statistics they were able to confirm that the items generated from the test work along an "ordered measurable dimension" of grammatical accuracy, and that this provides evidence for the validity of the tests developed. This ordered measurable dimension is described in terms of "task" difficulty in logits, the argument being that if these tasks (or rather items) can be ranked in terms of difficulty then it is possible to claim that a structured (presumably hierarchical) description of grammatical accuracy can be produced.

The work described above is useful in that it has shown one way in which the assessment of task difficulty may be approached. However, it would appear to be of limited value in direct oral testing as the "tasks" seem to be discrete items which are administered orally; there appears to be no good reason why they could not be
administered by paper and pencil methods to elicit the same language features.

It may also be noted that in band 1 of the scoring criteria it would appear that grammatical accuracy is not necessarily being measured, but the fluency with which the students can produce the required response. It is at least possible that this rating scale is confounding two potentially separate dimensions of language production, although this has not been investigated by the authors of the test.

These criticisms cannot be levelled at the work of Pienemann, Johnston and Brindley (1988), who used "spontaneous speech gathered in unstructured conversations" in their test and specifically related the scoring criteria to findings in second language acquisition. Pienemann et. al. (ibid., 220) argue that "second language learners follow a fixed order of stages in their acquisition of L₂ grammatical forms..." and developed a rating procedure which would allow profiling of learner development. The construction of the acquisition order was based on the analysis of transcripts of learner speech, originally in German and then in English, and was hypothesized to be governed by available speech processing devices beginning with non-linguistic devices and proceeding to language specific and then target language specific devices. This forms part of an implicational hierarchy which lead Pienemann et. al. (ibid., 225) to form the "Teachability
Hypothesis" which states that "the course of second language development cannot be altered by factors external to the learner." This is in line with work on second language acquisition conducted by Ellis (1987b).

Pienemann et. al. developed two observation forms using the acquisition hierarchy which they had observed in their data, so that raters could tick off occurrences of grammatical features listed. The features on the form were a sample from the entire range of features observed, and from this the acquisition of other features could be predicted from the hierarchical acquisition sequence described. In trialling their system it was found that the largest test method facet confounding results was that of rater, although they were able to account for this and provide reasonable initial validity coefficients for their method.

This work is noteworthy in the study of the assessment of accuracy in speech, for it shows that it is possible in theory and practice to use the analysis of spoken data to devise a rating scale (or observation form) which can then be used in live testing. Further, the final rating of the student can be used diagnostically to describe the stage of development which a given student has reached thus providing feedback to teachers which may affect future pedagogical input.
Section 2: Current criteria of accuracy in oral tests

Descriptors relating to accuracy will be reviewed in the following tests: The University of Cambridge Local Examinations Syndicate (UCLES) First Certificate in English (FCE), Certificate in Advanced English (CAE) and Certificate of Proficiency in English (CPE), the Australian Second Language Proficiency Ratings (ASLPR), the American Council for the Teaching of Foreign Languages (ACTFL) rating scale, and the International Language Testing Service (IELTS) band scales. The rating scales associated with these tests were selected for review because they represent the most widely used international tests which appear to rely heavily on accuracy as a rating criterion or in which, as is the case with the UCLES tests, a separate accuracy rating scale is used.

2.1 UCLES rating scales

In the UCLES suite of examinations it will be noticed that the band descriptors for accuracy tend to be similar across examinations, with the exception of the more recent CAE. In this latter rating scale lexis and grammar have been combined. As will be seen later in the Principal Components Analysis of data on twenty one students, there is evidence to suggest that vocabulary is actually a separate ratable component of accuracy, but that it is capable of being linked directly with levels of grammatical accuracy through discriminant analysis.
Whether it is preferable to keep the two scales apart or combine them may, therefore, be a matter of research into how many scales raters are capable of reliably handling within the live testing situation. This is something with which this research is not concerned.

The UCLES rating scales are given in full below.

**FCE: Grammatical Accuracy**

5  Use of a wide range of structures to deal with everyday topics and the expression of ideas and opinions without obtrusive inaccuracies.

4  Basic structures sound although there may be some inaccuracies when complex structures are attempted.

3  Basic structures sufficiently accurate to deal adequately with everyday topics though more complex structures may cause considerable difficulty.

2  Basic structures often inaccurate in everyday contexts. More complex structures not attempted or not intelligible.

1  Consistent distortion of basic structures.

0  No awareness of basic grammatical functions.

**FCE: Vocabulary Resource**

5  Appropriate range of vocabulary which allows for flexibility in dealing with everyday topics. Sufficient resources for expressing opinions, ideas and thoughts of a more independent, individual or specific nature.

4  Appropriate range of vocabulary to deal with everyday topics and sufficient to facilitate efforts to express views, ideas and thoughts of a more independent, individual or specific nature.

3  Range of vocabulary adequate for everyday topics, though experiences difficulty when required to expand on topics.
2 Vocabulary not adequate to deal with everyday topics.

1 Lack of vocabulary makes even basic communication very difficult.

0 Vocabulary not adequate even for minimal communication.

**CPE: Grammatical Accuracy**

5 Accurate use of a wide range of structures in dealing with everyday and discursive issues.

4 Use of a wide range of structures in dealing with everyday and discursive issues. Few errors when using complex structures and basic errors very rare.

3 Use of variety of structures in dealing with everyday and discursive issues. Some inaccuracies when using complex structures and a few noticeable basic errors.

2 Structures mainly accurate in everyday contexts but limited in range. Basic errors are noticeable.

1 No complex structures attempted and simple structures frequently inaccurate in everyday contexts.

0 Frequent basic errors

**CPE: Vocabulary Resource**

5 Range of vocabulary affords total flexibility and appropriate use in all contexts.

4 Use of vocabulary appropriate in most contexts. Sometimes needs to resort to paraphrase in more specific, abstract or academic contexts.

3 Range of vocabulary adequate for general topics. Often resorts to paraphrase or simplification in more specific, abstract or academic contexts.

2 Vocabulary resource adequate only for everyday topics. Seldom rises above the mundane, being dependent on paraphrase or simplification in specific, abstract or academic contexts.
Limited vocabulary resource makes performance, even in everyday contexts, inadequate.

Vocabulary too limited for communication at this level.

**CAE: Accuracy and Range**

7 - 8  Evidence of a wide range of structures and vocabulary in all contexts. Errors minimal in number and gravity.

5 - 6  Evidence of an appropriate range of structures and vocabulary; has the range needed to express intention. Errors few in number/minor in gravity. Message not impeded.

3 - 4  Fairly frequent errors and evidence that range of structures and vocabulary limits full expression of intent. Communication of the essential message is not prevented.

1 - 2  Frequent basic errors and limited range of structures and/or vocabulary impede communication of the essential message and constantly strain the listener.

0   Inadequate for assessment, even after prompting by the interlocutor.

In the CAE candidates who receive a 7 - 8 scores would be expected to pass the CPE, and those who receive 0 - 2 would be expected to fail FCE. Those who receive 3 - 4 would not receive the CAE, but would be expected to pass the FCE, whilst bands 5 - 6 are considered good CAE candidates. It should be noted that the CAE scale differs from others of its kind (with the one exception of the oral rating scale for the International General Certificate of Secondary Education English as a Second Language Examination) in that the descriptors are provided for a range of bands rather than individual
bands. This is to allow the examiner to match a student with a descriptor and then allow him/her to decide whether it is a superior or inferior performance within the band range. As there is no evidence available on which to judge whether or not this approach leads to higher reliability, or whether the method is valid, it is not possible to assess the scale with regard to this property.

With regard to the FCE and CPE scales, it can be seen that some attempt has been made to discriminate between the levels of candidate expected to attempt each of the examinations. This is most clear in the CPE Grammatical Accuracy bands 3 - 5, where reference is made to "discursive issues", not included in the FCE scale. However, the other band descriptors rely on concepts such as "everyday issues" and "complex" and "basic" structures which remain undefined. This is a basic problem with the scale. As Pienemann et. al. (1988: 219) point out, what is viewed as simple or complex in terms of structure may not be simple or complex for language learners, and such terms must always be defined in relation to a theory of language acquisition with empirical underpinning.

In the CAE scale the notion of error gravity is introduced (see pages 192 - 193), but no indication is given of which errors are more serious than others. The difference between bands 7 - 8 and bands 5 - 6 appears to depend upon the difference in interpretation between "wide" range of structures and "appropriate" range of
structures, and a "minimal" number of errors compared with "few" errors. Again, this terminology is not defined. Similarly, in bands 5 - 6 the message is not "impeded" whilst in bands 3 - 4 the "essential" message is not "prevented". It would appear that all terminology is in need of precise definition.

However, it could be countered that examiners' awareness of which students should be placed within each category is determined by the video cassettes provided for rater training. These are provided for rater training under the guidance of local team leaders around the world, but even if evidence were presented to show that this increases the reliability of rating it does not mean that the scale is automatically valid as it stands without the definition of key terms. Indeed, it has been argued that once rater training takes place there would be no need for band descriptors at all, as standards are imposed through the training process (see Chapter 2.1.3). Matthews (1990), for example, conducted an experiment with the descriptors for the International General Certificate of Secondary Education (IGCSE) in which teachers were asked to place jumbled descriptors in order according to the original proficiency level described in the scale. This proved to be impossible. This reflects on the validity of the descriptors, not upon their reliable use after raters have been trained.

It would therefore appear to be the case that the principle expressed by Lado (1961), Baker (1989) and
Hughes (1989) that scale descriptors should be as explicit as possible have not been met in these scales. However, it will be argued that this is in fact the case with other descriptors in rating scales which are currently in use.

2.2 The ASLPR

The ASLPR speaking descriptors are much more detailed than many others currently in use, allowing raters to look for specific aspects of accuracy in candidate performance. The phrases which relate to accuracy have been extracted from the descriptors and are presented below. The comments are provided as notes for the rater on the scales (Ingram and Wylie, 1985).

S:0+ Vocabulary limited to that necessary to express simple elementary needs and basic courtesy formulae. Syntax is fragmented, inflections and word endings frequently omitted, confused or distorted...

Comment: Distortion of word endings may involve omission, addition or substitution of phonemes or allophones.

S:1- Most utterances are telegraphic and word endings (both inflectional and non-inflectional) are often omitted, confused or distorted.

S:1 ...fractured sentence structure and other grammatical errors are frequent.... Misunderstandings frequently arise from limited vocabulary and grammar....Little precision in information conveyed owing to tentative state of grammatical development and little or no use of modifiers.

Comment: Modifying devices or modifiers are those forms (eg., verb forms, adjectives, adverbs, phrases,
clauses, etc.) that are used to modify and give precision to the expression of thought.

S:1+ The commoner tense forms occur but errors are frequent in formation and selection. Can use question forms. While basic word order is established, errors still occur in more complex patterns....Aware of basic cohesive features (eg., pronouns, verb inflections), but many are unreliable.

S:2 ...can usually handle elementary constructions quite accurately but does not have thorough or confident control of the grammar especially in longer constructions.

S:3 ...control of grammar good; able to convey meaning precisely in reasonably complex sentences or by using, with reasonable accuracy, a wide range of modification devices.

S:4 ...errors of grammar are quite rare and unsystematic and can usually be corrected in retrospect...

Comment: Grammatical development is now more or less complete though "slips" or errors of performance may still occur. The learner can, however, usually correct such errors if he becomes conscious of them.

S:5 Has complete...accuracy...in the language such that the learner's speech on all levels is fully accepted by such native speakers in all its features...

The band descriptors are more specific than are usually found in rating scales, but it may be observed that only modification devices are specified in detail in S:1. The other descriptors rely on the familiar concepts of basic versus complex structures and relative degree of accuracy of production.

The scale is more specific at the lower range of band descriptors, as Ingram (1982) states. Unlike the UCLES scales it is global in nature, but would appear to
have been constructed along similar lines: the authors wrote the descriptors on the basis of their view of what constitutes proficiency at any given level from their own intuition, experience and studies in psycholinguistics. This interpretation would seem to be correct, as Ingram (1982: 10) comments that L2 research has not discovered any universal pattern in language development, but that scales implicitly rely on the notion of such a pattern. This view will later be called into question, although it is interesting to note that Ingram (1985a: 4) himself claims:

"We have noted that rating scales such as the ASLPR are developmental in structure, i.e., the progression of behavioural descriptions attempts to reflect the way in which a second language develops from zero to native-like. Thus the scale is not arbitrary but related to the universal development schedule and a learner's rating indicates the stage of development he has reached in the developmental schedule."

Although the ASLPR has been researched to some degree, the differences between the 1982 and 1985 claims demonstrate that the so-called developmental structure of the rating scale is not actually based upon any empirical research into a "schedule" of language acquisition. This, of course, is not to say that the scale itself might not represent such a schedule in its general form, but that it is not specific enough in detail to describe the schedule or indeed to be validated against any schedule.
Nor is this to say that the scale could not be used for research into such a schedule if the band descriptors were expanded, made more specific, and appropriate research conducted.

2.3 The ACTFL rating scale

The ACTFL rating scale is another holistic criterial approach (Hughes, 1989: 103 - 104; 110) in which candidates are matched to a descriptor on the basis of their performance on a task. The ASLPR as a holistic scale does not focus on accuracy specifically, but has accuracy as part of its descriptors. In contrast, the ACTFL scale is seen as placing accuracy in prime position (Kramsch, 1986b). This view is, to some extent, unfair. Writers who have supported the Proficiency Movement which uses the ACTFL scale as a guide to curriculum development may have stressed the grammatical accuracy aspect of the scale, but the scale itself merely includes references to accuracy among other aspects of performance. Kramsch argues that the level of "abstraction" - the non-specificity of the criteria described in the ACTFL descriptors - makes the scale uninterpretable. Her criticisms of the emphasis on grammatical accuracy in particular are more related to the format of the elicitation techniques, which are geared towards eliciting particular grammatical structures, than to the notion of assessing accuracy within an oral test per se.

Although Kramsch's criticism of the non-specificity
is essentially correct, the ACTFL scale does still represent a trend towards greater specificity in rating scales, notably at the Advanced Plus level.

Van Patten (1986) also criticises the position given to grammatical accuracy in the ACTFL scale. Although the ACTFL scale is certainly one of the most researched scoring instruments and has produced much valuable research (Skehan, 1991: 9 - 12), Van Patten is probably correct in his argument that in the development of the ACTFL scale (based upon its predecessors, the FSI and the ILR) no database of information was used in the drafting of the descriptors. He correctly points out that in defending the validity of rating scales expressions such as "experience suggests..." and "experiential data..." are frequent, while actual empirical evidence is sparse.

The point is that validity, if it is to be understood in terms of the relationship of the descriptors to actual language production or sequence of acquisition, was not considered at the outset of the construction of the scale in anything but an intuitive way, and we must therefore remain extremely cautious in interpreting the descriptors as if they represented psychological or linguistic reality, even if the scale is widely accepted and shown to be reliable after rater training. Similar criticisms can be levelled at all rating scales currently in existence, not only the ACTFL.

References to accuracy in the ACTFL rating scale are here extracted at the various levels of proficiency.
described (ACTFL, 1986).

**Novice-Low**

[As the candidate cannot function in the language, no mention is made of accuracy.]

**Novice-Mid**

Syntax is fragmented, inflections and word endings frequently omitted, confused or distorted...

**Novice-High**

Can ask questions or make statements with reasonably accuracy only where this involves short memorised utterances or formulae. Most utterances are telegraphic and word endings are often omitted, confused, or distorted.

**Intermediate-Low**

Is able to formulate some questions with limited constructions and much inaccuracy. Almost every utterance contains fractured syntax and other grammatical errors.... Misunderstandings frequently arise from limited vocabulary and grammar.... Little precision in information conveyed owing to tentative state of grammatical development and little or no use of modifiers.

**Intermediate-Mid**

Some evidence of grammatical accuracy in basic constructions, for example, subject-verb agreement, noun-adjective agreement, some notion of inflection.

**Intermediate-High**

The commoner tense forms occur but errors are frequent in formation and selection. Can use most question forms. While some word order is established, errors still occur in more complex patterns. Cannot sustain coherent structures in longer utterances or unfamiliar situations.... Aware of basic cohesive features such as pronouns and verb inflections, but many are unreliable, especially if less immediate in reference.

**Advanced**

Can usually handle elementary constructions quite accurately but does not have thorough or confident
control of grammar.

Advanced Plus

Generally strong in either grammar or vocabulary, but not in both.... Areas of weakness range from simple constructions such as plurals, articles, prepositions, and negatives to more complex structures such as tense usage, passive constructions, word order, and relative clauses.

Superior

Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively.... Control of grammar good; errors virtually never interfere with understanding and disturb the native speaker.

The resemblance of the ACTFL and the ASLPR is most striking, with only minor differences in wording, the ASLPR having been generated from drafts of the ACTFL scale. As such, similar comments apply: the ACTFL scale could be used as the basis for further research and development if the descriptors were made more precise and directly related to studies in second language acquisition. At the present time, it remains too vague to be used in such a way (Fulcher, 1989).

2.4 The IELTS rating scale

The IELTS oral rating scale is the successor to the ELTS rating scale, the latter having been discussed in Fulcher (1987), and used in rating the 21 students whose interviews formed the initial database for rating scale development in this study. Developed in conjunction with Ingram, it would be surprising if aspects of the IELTS
scale did not resemble the ASLPR. References to accuracy in the nine bands of the IELTS have been extracted, and are presented below.

Band 9

[No reference to accuracy]

Band 8

Non-systematic errors in grammar (and) vocabulary.

Band 7

Errors in vocabulary and structure may occur without inhibiting communication. Communicates readily and fairly precisely using complex sentence forms and a wide range of modifiers, connectives and cohesive features.

Band 6

Can use complex sentence forms and a wide range of modifiers, connectives and cohesive features to convey most meanings fairly precisely though errors in grammar and vocabulary may occur and occasionally interfere with communication. Is generally able to use circumlocution to cover gaps in vocabulary and structure.

Band 5

Errors in structure and vocabulary may interfere...generally making use of relevant connectives.... Has some ability to use complex sentence forms and modifiers.

Band 4

Can use common question forms to elicit information (though not necessarily with correct word order). Has control of basic sentence forms but longer utterances tend to break down. Can link simple sentences using the most frequently occurring connectives. Errors in grammar and vocabulary are frequent.... Tentative use of modifiers...

Band 3

Basic sentence forms appear to be used though grammatical errors are numerous except in memorized utterances. Essentially no ability to link sentences of use modifiers.
Bands 2 - 0

[No specific reference to accuracy as, it must be assumed, students at these levels do not provide evidence that any grammar or vocabulary, short of memorized words or phrases, can be used.]

The "developmental schedule" (Ingram, 1985a) of the ASLPR does seem to be reflected in this rating scale, although the level of specificity is more general than the ASLPR rating scale. Information regarding how the scale was constructed is provided in Alderson (1991a: 81). This involved: the use of expert judges who commented on the scales; getting experienced markers to place samples of speaking against each band and then decide which "key features" led them to their decision; collection of feedback after the scale had been put into use. The latter is different from the other two methods of scale construction in that it is a post-hoc assessment of validity. The second of these methods is the closest to what is being presented here as an appropriate approach to the inclusion of considerations of validity at the design phase of the band scales, and it would be interesting to have more details of how, and to what extent, these markers agreed on the distinguishing features of performances which discriminated between bands on the scale. However, the main issue appears to have been the use of quantifiers. The problem could have been the pre-existence of the nine bands and the need to describe them, when the number of bands that can reliably
be distinguished does not seem to be that high (Pollitt, 1991: 90), and in this research only four bands could be distinguished and described from the data available.

Section 3: Specificity in Rating Scales

It has been seen that the degree of specificity or of abstraction in band descriptors within a rating scale is an issue of some importance, and one which was raised consistently by the raters used in the present study (see Appendix 10.1 and pages 375 - 376). When using a rating scale in a live oral interview there is the very real problem of deciding how much information the rater can process: the more general the scale and the shorter the descriptors the easier it is to use. The emphasis must then be placed on rater training using video/audio interviews to establish standards by linking particular performances with band descriptors. The more specific the band descriptors, and hence the larger the physical document being used, the harder it is to operate the system in a live testing situation. As Porter (1991: 32 - 33) says:

"...the more complex the description gets, the less our brains are able to grasp it in its entirety and the less it means to us."

Porter argues that the descriptors which are useful are those which "leave a considerable amount out", and that selection of salient features for inclusion will be
an important part of scale development. This problem
should be borne in mind when constructing a more specific
rating scale. In this project, the appropriacy of the
amount of detail included in the rating scales,
substantially more than other rating scales to date, was
assessed by asking raters to retrospectively comment on
their uses of the scales and talk about any problems
which they encountered.

Section 4: Developing a rating scale for accuracy

In the initial stages of developing a rating scale
for accuracy a similar procedure to that used in the
development of the Fluency scale was envisaged, beginning
with an analysis of the same twenty one transcripts and
tapes of oral interviews and the isolation of categories
of error which may discriminate between student ability
levels. A sample coded transcript is contained in
Appendix 5.21. In total, nine categories were isolated,
but many of these categories did not contain enough data
to be able to progress immediately to discriminant
analysis.

It became obvious that some of the categories would
have to be combined in order to make progress. It was
initially suspected from the nature of the descriptive
categories developed that the data could be reduced to
two macro-categories termed "grammatical accuracy" and
"lexical accuracy", but in order to investigate if this
distinction had any basis in the actual observations
Principal Components Analysis was employed. The categories were then combined and discriminant analysis used to estimate the ability of the new categories to discriminate between students at different levels.

Section 5: Analysis of categories

Examples of inaccuracy in language use identified from the transcripts were tallied. These were then grouped into the nine categories described below. The definitions of the categories were adapted and expanded to suit this data from Burt and Dulay (1975). If a grammatical or lexical item was used incorrectly it was counted as an error even if corrected in later speech, and if an item did not occur in an obligatory slot within the utterance it was also counted as an error by omission. In using this procedure it should also be noted that there is a certain amount of indeterminacy in categorising aspects of the data. It should be borne in mind that (a) isolating an "error", "mistake" or "slip" in spoken data does not mean that the researcher can automatically know under which of the headings to classify it; this has to be inferred from the context; (b) the error may have been the result of lack of knowledge (ie., the structure has not been acquired), difficulty with processing language or cognitive overload; once again it is hard to isolate causes; and (c) once an error has been identified it is difficult to classify the error in many cases. For example, if the
researcher is faced with the utterance: "...and he passed from the other department..." are we dealing with a prepositional error ("by" rather than "from") or a lexical error (misselection of "pass" when some other verb was required)? In this case, the error would have been classified as the former on the grounds that the L₁ of the students used to construct the database was Greek, and the use of a few prepositions in Greek to do the work of the many in English leads to typical errors of which this may be one. However, it cannot be denied that there is indeterminacy here, but this is, and will remain for the foreseeable future, a problem with all investigations of this nature. It would certainly seem to be a problem with the morpheme studies discussed later, although this has not been tackled to any extent by researchers in that area. A coded transcript is included in Appendix 5.21 for the reader to examine.

The initial nine categories of accuracy phenomena are listed and described below.

**Category One: Word Order**

This category included any examples of simple or compound declarative sentences where the order of constituents was incorrect, as in Example 5.1.

**Example 5.1.**

B> because er: is different Cyprus from the London people..."
In context the student intended to express the belief that people in Cyprus are different from those in London.

**Category Two: Pronouns, relative pronouns and dummy pronouns**

Burt and Dulay only consider pronouns in nominative or accusative case in simple pronoun subjects or objects, as in Example 5.2:

Example 5.2.
B> ...her father don't er: if if he if he like to live in the house he must paid er the rent...

In this case the student wishes to refer to girls who remain in their parents' home after completing school paying rent. The correct pronoun required is "she", but the incorrect male pronoun is continuously supplied.

Dummy pronouns relate to the use of "it" and "there" to refer to a nominal group which will lexicalise (give content to) the pronoun after the next verbal group in the utterance.

Example 5.3.
B> ...er I don't know if there are in: in towns but: in villages er: a lot of illiterates people..."

In Example 5.3 we would expect the student to say "there are a lot of illiterate people", but the student does not seem to have acquired this use of the pronominal
Relative pronouns were found to cause more problems in speech than pronouns in the other two categories, mainly through omission, as in Example 5.4.

Example 5.4.
B> ...then we have we have r is the yield er: of the material...

In Example 5.4 the meaning of the utterance would have been clearer had the student said "...we have r, which is the yield (point) of the material." In Example 5.5, on the other hand, the incorrect "where" should have been "which".

Example 5.5.
B> ...where the load is fourteen kilonewtons: where is the elastic point....

Category Three: Tense

This category combined many of the individual morphemes which are often studied separately, including present simple (and third person -s), past regular and irregular, the present progressive, perfect auxiliary and past participle. Examples 5.6 and 5.7 illustrate the type of errors which were categorised under Tense.

Example 5.6.
B> ...until then I have been work in Nicosia General Hospital..."
Example 5.7.
B> I am in the Political Affairs Division for the last five years...

Category Four: Lexical Errors

It was discovered that many of the errors in the transcripts were of a lexical nature. These included incorrect word formation, lexical problems with time reference, misselection of an appropriate modal verb, and confusion over the use of word pairs such as make/do, get/take and say/talk. Examples 5.8 to 5.10 illustrate the types of errors placed in this category.

Example 5.8.
B> ...and er::: the adult literary is raising...

Example 5.9.
B> ...er the course was only until six months but I stayed for nine months...

Example 5.10.
B> ...we don't have such a laboratory to make this but er...

Category Five: Omission of subject, verb phrase, or both

On a number of occasions students began an utterance with the verb phrase, omitting the whole of the subject of the sentence (whether this is a pronoun or noun phrase), leaving out the verb completely, or simply leaving both out. In Example 5.11 the student omitted "you" before the phrase "have to bear in mind", whilst in
Example 5.12 the utterance could be made to sound more natural by supplying a pronoun and the verb "to be", as in: "because I am er a sergeant in the army..."

Example 5.11.
B> Yes [yes A> (laughter)] B> have to bear in mind...

Example 5.12.
B> I have my office right and er: because er sergeant in the army I have some responsibilities..."

Category Six: Errors in prepositional phrases

Selection of the correct preposition for a phrase was a major problem for many students, even though this is a category of errors which does not seem to have been studied extensively.

Example 5.13.
B> ...they are first checked from their quality and then registered here...

Category 7: Formation of plural nouns (and this/these)

Long /:iz/, short /s/, /z/ and irregular plurals were taken into account in this category.

Example 5.14.
B> ...to come back in Cyprus to make some lectures to other... [others]

Example 5.15.
B> ...fifteen thousand er women went to prison: er: er: er:::

195
Category Eight: Articles

In the morpheme studies "a" and "the" are combined under the general heading of articles and are not treated separately.

Example 5.16.
B> ...(NAME) is a small college for the English (language)...

Example 5.17.
B> ...and will do: er: foundation course...

Category Nine: Miscellaneous

Tallies of errors were recorded in a number of areas which could not be treated as separate categories because of the extremely small number of examples from the data. These included errors with the formation of comparative and superlative adjectives, conditional tenses, genitives, use of connectors, and use of the gerund or infinitive constructions, as exemplified in Examples 5.18 - 5.22.

Example 5.18.
B> er er it's er more easy with er law students studying in England...

Example 5.19.
B> ...if if I will go to London I will live with English family...
Example 5.20.
B> ...the ladies er leave home: er things: to to marry then go out A> hm B> of er his father home...

Example 5.21.
B> ...we are er: to twelve at and er: in Nicosia General Hospital...

Example 5.22.
B> ...we can see that the load increases and then it start it’s starting er decreasing...

Section 6: Error Gravity

Prior to beginning the analysis of the categories of error described above, the issue of error gravity was investigated. It was initially suspected that work in this area might provide a theoretical framework within which categories could be combined for analysis, or provide insights regarding the extent to which certain categories should be stressed at particular levels within a rating scale.

Work in the area of error gravity such as that of Burt and Kiparsky (1972) and Burt (1975) has concluded that errors in speech fall into two main categories: global errors which affect the communicative intention of the utterance by breaking up the grammatical meaning of whole chunks of language, and local errors which are limited in the effect on the utterance and so do not cause problems with communication.

It may therefore be suspected that with regard to the nine categories described above, Category 1 (word
order), Category 5 (omission of noun or verb phrases), and to some extent Category 2 (pronouns, only when the relative clause is affected), would constitute the most serious errors in terms of the communicative value of the utterance. Errors in all other categories in the study would, it was thought, tend to have a locally disruptive effect but would probably not tend lead to a total breakdown in communication.

As will be seen in what follows, the use of these results from the error gravity studies in developing a rating scale for accuracy would not prove useful. No matter how intuitively satisfying a scale which takes error gravity into account would be (most problems with word order in lower ability candidates, least in higher ability candidates etc.), the descriptors could not be empirically justified. For example, the present analysis found that such global errors only tend to manifest themselves in the speech of students in the middle bands of a rating scale, as candidates of lower ability did not attempt to use extended utterances in which such errors could occur. A rating scale must therefore take into account what actually occurs in speech at various levels of ability.
Section 7: Initial investigation into category reduction

Treating each observation of an error as an "error score" on each of the categories described above, total error scores were correlated. It was expected that there would be evidence to support the supposition that two categories: grammatical accuracy and lexical accuracy would be enough to account for the data. The correlation matrix presented in Table 5.1 in the Appendix to Chapter 5 was submitted to Principal Components Analysis using Varimax Rotation in order to see if there was any principle upon which the categories could in fact be combined before proceeding to use discriminant analysis. It was hypothesized that there would be two factors, lexical accuracy consisting of categories four and six, and the other grammatical accuracy consisting of all other categories.

Table 5.2 in the Appendix to Chapter Five shows that although two factors can be isolated from this correlation matrix, it is categories 2 (pronouns) and 5 (omission of noun or verb phrases) that load highly on factor two. Factor 2 is therefore characterised by students who do not omit noun or verb phrases but who tend to make errors in the use of pronouns, or students who use pronouns accurately but do tend to omit noun or verb phrases.

Referring to the original transcript data it was discovered that of the 41 pronominal errors in the tally, 24 came from candidates who had received a band score of
6 on an ELTS interview. Five of these were examples of the misuse or omission of the appropriate object pronoun. All other examples occurred as a result of the more complex sentence structures being attempted by these candidates.

Category 2 was omitted from the discriminant analysis and reference to pronominal error included only in the final scale in band 2 as a partial criterion. Category 5 was treated in the same way, and reference to omission of sentence constituents included only in bands 1 and 2.

It had, however, been judged essential not to produce an over complex rating scale by mixing information on lexical and grammatical accuracy, but to present these side by side on the rating scale. This would make the rating scale easier for the raters to read, although there was no indication that lexical and grammatical accuracy were indeed separate components. Categories four and six were therefore used as a lexis category despite the finding from the Principle Components analysis that they did not load on a separate factor from other categories used in the study. This decision allows the use of discriminant analysis, but the separation was not used to create a separate rating scale for lexis, as this would have been unjustified given the findings of the PCA. The separation did not, therefore, have any direct effect on the construction of the Accuracy rating scale.
Section 8: **Discriminant Analysis**

Before conducting the discriminant analysis the total number of errors for each of the twenty one students in the two categories was adjusted for the amount of language produced during the interview. This was done because many of the candidates who talked at length produced more errors than candidates who did not produce so much language. However, error density was not as great. The adjustment procedure used was the same as that described by Shohamy, Shmueli and Gordon (1991), namely to divide the frequency of errors by the total number of words produced.

This adjustment in itself introduces another element into the concept of the accuracy scale, that of the amount of language produced, which will be considered later (see pages 204 - 206).

As in the development of the Fluency scale, discriminant analysis was used to determine whether or not the observations made were capable of discriminating between ability levels. In this analysis the two lower ELTS bands (4 - 5) and the upper two bands (8 - 9) were collapsed into one band each, producing a possible 4 band scale if the results proved to be significant. This was done on the grounds that initial group means on each of the categories for ELTS bands 4 and 5 were very similar, indicating that there was no difference between them and hence there was no good reason to maintain them as separate in the process of constructing the Accuracy
rating scale.

The multivariate test results are presented in Table 5.3 in the Appendix to Chapter 5, and it can be seen that Wilks' Lambda is significant at \( p = .003 \) indicating that the two categories used discriminate well between the students. Second and subsequent discriminant functions from the analysis were not significant.

In order to test the accuracy of the predictions which could be made from the observed tally of errors, the actual band scores which were awarded to candidates may be compared with the band scores they would be awarded if only accuracy on the two categories were taken into account. This is exactly the same procedure used in the production of the Fluency scale (see Figure 4.1), and Figure 5.1 presents the predictions for accuracy.

It can be seen in Figure 5.1 that of the number of students actually awarded a band 1 (originally ELTS bands 4 - 5), 2 would be awarded a band 1 and 5 a band 2 (originally ELTS band 6) on predictions from this analysis. Of those awarded a band 2, six of the eight students would again be awarded a band 2, the other two students being awarded a band 1. However, it is in the band 3 (initially an ELTS band 7) where the maximum error could occur. Of the four students who received a band 3, 3 would receive a band 1. Of the candidates who were awarded a band 4 (originally ELTS bands 8 - 9), one would still receive a band 4 and one would receive a band 3.
Fig. 5.1. The relationship between actual band scores and predicted band scores for accuracy.

<table>
<thead>
<tr>
<th>Band Awarded</th>
<th>Predicted Band</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>11</td>
</tr>
</tbody>
</table>

The amount of error involved here, especially with band 3 students, could indicate one of two possibilities: either the assessment of accuracy in live oral interviews involves a great deal of inherent indeterminacy (see page 136), or secondly that other aspects of production are given greater weight in awarding global band scores. The second of these possibilities could be the case in the light of the accuracy of the predictions made from the Fluency scale developed in this study in Figure 4.1, in which only one student would have been misplaced by one band. That is, the concept of fluency is closer to what raters actually rate in live interviews. However, more light will be thrown upon this issue in the validity study. If the assessment of accuracy is indeterminate even when raters are asked to concentrate upon it, then the validity coefficient for the Accuracy scale would not be expected to be very high.

In the preceding discussion it can be seen that the
development of the Accuracy rating scale faced difficulties from the very beginning, and the fact that prediction from isolated linguistic phenomena which are not "high inference" but observational categories do not result in high prediction rates to band scores may suggest that the meaningfulness of the construct of Accuracy may be called into question. This is to be compared with the development of the Fluency scale where the descriptions, however high inference in nature, led to what appeared to be a very stable construct at least at the scale design phase. It is in the validation study that the implications of these design problems may be discussed and assessed.

Section 9: Amount of speech produced

It was stated that the number of errors in certain categories related to the amount of speech produced. It was observed, for example, that lower ability students did not make global errors for the simple reason that they do not often produce strings of speech which would allow the error to occur. Looking at the transcripts, it appeared that a large number of turns produced by lower ability students were single words or fillers, whilst with higher ability students occurrences of single word turns also included backchanneling. The remaining examples of single word turns were either single lexical items, or "yes" and "no".

Table 5.4 in the Appendix to Chapter 5 shows the
total number of turns taken by each of the twenty one students, the number of single word turns as a percentage of the total number of turns, the percentage of fillers in single word turns, and the percentage of backchannels as single word turns. From these figures it was hypothesized that the number of single turns and the number of backchannels would also discriminate well between lower and higher ability students. For the purposes of investigating the number of single word utterances and the phenomenon of backchanneling, the original bands were collapsed to form just three: ELTS 4 - 5 forming band 1 (as in the accuracy analysis) with band 6 forming band 2 on its own, and bands 7 - 9 forming band three, as an initial investigation of means indicated that there was no apparent difference between these bands. This was then submitted to discriminant analysis. The multivariate results are presented in Table 5.5 in the Appendix to Chapter 5.

The results are significant at \( p = .002 \), and by analysing Table 5.4 we may see the tendency for lower ability students to use more single word utterances than higher ability students, and the tendency for higher ability students to use more backchanneling than lower ability students.

The results regarding the number of single word utterances were used in the construction of the Accuracy scale in order to take into account the relationship between number and type of errors made and the amount of
language produced. The information regarding backchanneling was used to improve the descriptors in the Fluency scale, as backchanneling is not related to the accuracy of an utterance, but to the perception of fluency in maintaining conversational flow.

Section 10: Breadth of lexis

From the data it was observed that students in the middle bands of a rating scale tended to make more errors of the kinds described in Category 4 (lexical errors). In order to interpret this for the development of the lexical accuracy descriptors, it was necessary to look at any factor which may have explained this. It was hypothesized that on a four band scale, band 2 candidates would begin to use a much greater range of vocabulary and in this attempt to be more informative make more errors. This would imply that band 1 students had a restricted lexical range.

In order to investigate this the total number of words uttered in each interview was counted, the number of unique words was calculated (words which were used only once by the student), and the number of word forms (the word "you" may occur ten times, but is only counted as one form). The number of unique words was also calculated as a percentage of the total number of words, and the total number of words divided by the number of word forms was calculated as an estimate of lexical range.
If any of these estimates of the students’ use of lexis could discriminate between band 1 and band 2 students, this could be included as a factor in the Accuracy rating scale. The results of the analysis are presented in Table 5.6 in the Appendix to Chapter 5, and the means for total number of words, unique words and number of word forms are contained in Table 5.7. In Table 5.7 it may be seen that the change in pattern occurs between Band 1 and Band 2. These factors were thus taken into account when describing the types of lexical errors which were more likely to occur at band 2 than band 1.

It may be added here that this research very quickly pointed up the fact that oral rating scales cannot be "linear" in the way which a-priori scales are. That is, more errors may occur in the speech of more able students because they are in the process of acquiring more language and experimenting with it. Similar results were discovered by Meisel (1980) who noticed increasing "wrong use" forms as earlier systems were extended in the direction of the second language system, and called this transitory interlanguage "elaborative simplification" in contrast to the previous "restrictive simplification." Thus, a data-driven approach would appear to be superior to a-priori scale development, the linear nature of which does not take into account how students acquire language and actually speak.
Section 11: Producing Accuracy rating scale descriptors

Having produced a number of observations which succeeded in discriminating between students in 4 bands it is possible to return to the original observations on the transcripts of the oral interviews as described in the nine categories of error identified and use these to construct verbal descriptors which represent levels of ability. As lexical and grammatical accuracy were considered to be separate but highly related aspects of accuracy, which nevertheless generated four separate bands, the rating scale maintained this distinction in format.

Table 5.8 contains simple correlations between band and the 9 categories in the accuracy analysis, and Tables 5.9 to 5.17 provide the mean scores on each of the categories and band scores. These means are plotted in Figures 5.2 to 5.10 on pages 209 to 213. (It must be remembered when reading Tables 5.9 to 5.17 that these means are not averages of raw counts of errors in the categories. Such data in itself would be meaningless as the variable of amount of language produced would mask differences in error rate, as argued on pages 204 - 206. These means were produced after correction for amount of language produced).

It will be noticed from Table 5.8 that none of the correlations are very high, indicating the lack of a clear linear relationship between category and band. However, the Table 5.3 does show that the categories help
Fig. 5.2. Category 1 means

Fig. 5.3 Category 2 means
Fig. 5.4 Category 3 means

Fig. 5.5 Category 4 means
Fig. 5.6 Category 5 means

Fig. 5.7 Category 6 means
Fig. 5.8 Category 7 means

Fig. 5.9 Category 8 means
to discriminate between bands. The reason for these correlations appears to be that many of the observations are not linear across bands, as was seen in the process of constructing the Fluency rating scale. One example of how a band descriptor was constructed will be given with reference to Table 5.10 and Figure 5.3 for category 2, pronominal reference. The number of reference errors made by Band 1 students was less than both band Band 2 and Band 3 students. Only Band 4 students had less errors (in fact no errors of pronominal reference were observed from Band 4 students). This relates directly to the finding (Section 9 above) that Band 1 students produced less speech than students in other bands, and that their utterances were often less complex. As ability increases more errors are made as the students experiment with a wider range of grammatical structures requiring cohesive
reference, and as ability increases further (Bands 3 and 4) more accuracy in the use of cohesive devices is achieved. Hence, we see the inverted "U" pattern in Figure 5.3 which explains the .09 correlation coefficient between band and category 2 in Table 5.10.

In constructing the band descriptors for the Accuracy rating scale pronominal reference was not a feature of the Band 1 descriptor, as utterances were short and simple. The Band 2 descriptor specifically highlights the expected increase in the length of utterance and the complexity of language which will be produced, which will also lead to more errors in pronominal reference. After Band 2, the number of errors made begin to decline with increasing ability. This type of error was therefore used to characterise a Band 2 student.

It was assumed, as in the construction of the Fluency scale, that raters would tend to avoid extreme categories, and that a 4 band scale would in effect mean that bands 1 and 4 would not be used. The rating scale developed therefore contains 6 bands. Band 0 was merely described as performance inferior to band 1, and band 5 as performance superior to band 4. This, it was assumed, would encourage raters to use the entire range of bands 1 to 4 which can be empirically justified.

In the following rating scale, comments on the relationship between the descriptors and the data from which they were generated is enclosed in square brackets.
In the rating scales used by the raters, accuracy of grammar and accuracy of lexis were presented side by side rather than one under the other, as the descriptors were intended to be used in parallel to award one accuracy score.

**Accuracy of Grammar**

**Band 0**

Candidates in band 0 do not reach the required standard to be placed in band 1.

**Band 1**

Most band 1 candidates have trouble with forming plural nouns, and may also misuse this/these and much/many [Category 7]. Errors occur in the use of the genitive forms, conditional tenses, and the use of comparatives and superlatives (if attempted) [Category 9: frequency counts show that band 1 candidates make most errors in the use of these structures]. Frequent mistakes in the use of the article occur [Category 8]. Sometimes verb and noun phrases may be omitted from the sentence entirely [Category 5]. Verb errors typically include: past simple when narrating, present simple and present perfect; this may very often involve using some other tense than the one required [Category 3]. Band 1 candidates tend not to use many connectors other than "and", but often misuse other connectors if attempted [Category 9]. As utterances tend to be short and simple, there are few errors with word order [drawn from evidence regarding length of utterance and the finding that errors from Category 1 tend not to occur with lower ability students].

When a band 1 candidate is monitoring his/her production and realises that an error has been made, he/she may attempt to correct it, but the reformulation is as likely to be inaccurate as correct. [In the development of the Fluency rating scale, Category 6 dealt with hesitation during grammatical and lexical repair. It was discovered that whilst there was a difference between students regarding time spent on reformulating utterances, there was also a difference regarding the success of the reformulation. Whilst the time spent in reformulation was used in the construction of the Fluency rating scale, it was considered that the accuracy rate of the reformulation should be used in the definition of the Accuracy scale].
Band 2

Band 2 candidates may, like band 1 candidates, have trouble forming plural nouns [Category 7] and may also omit whole noun or verb phrases from time to time [Category 5]. Frequent mistakes in the use of articles also occur [Category 8].

In verb forms, however, band 2 candidates have few problems with the use of past tense, although they will misuse the present simple, present continuous and present perfect tenses [Category 3].

Band 2 candidates are most easily identified by their attempts to use more complex grammatical structures in communication, although this does not always result in greater accuracy. [This relates to the finding that band 2 candidates tended to make more errors due to the increasing length and complexity of utterances]. They may often make mistakes in the use of relative pronouns, relative clauses and pronominal reference [Category 2].

When a band 2 candidate is monitoring his/her own speech and notices that an error has been made, he/she will attempt to correct it. Unlike a band 1 candidate, reformulations will tend to be accurate.

Band 3

Although the frequency of extended contribution of band 3 candidates to the conversation is higher than band 2 candidates, only occasional errors can be noticed in word order or clause structure [Category 1: word order errors occurred in the speech of band 2 candidates because of the longer contributions; at band 3 the length of utterances increases further, but the number of errors decreases]. Mistakes do occur from time to time with the formation of plural nouns [Category 7], and the misselection of an appropriate (or zero) article [Category 8]. In verb forms only mistakes in the use of the present simple tense are noticeable [Category 3]. However, mistakes in other areas of grammar are very rare and non-systematic. When a mistake does occur the candidate’s self-correction is almost always successful.

Band 4

Band 4 candidates rarely make mistakes in their use of grammar. When they do these are always non-systematic, typically including minor mistakes in word order [Category 1] or the misuse of an article [Category 8].
Band 5
Candidates in band 5 reach a standard higher than that described in band 4.

Accuracy of Lexis

Band 0
Candidates in band 0 do not reach the required standard to be placed in band 1.

Band 1
Most lexical errors are concerned with the use of modal verbs with incorrect colligation (structure of the modal clause). "Must" is a common problem of usage within this category. Errors in the use of prepositions are extremely noticeable, but few other lexico-grammatical errors occur, as the typical band 1 candidate gives the impression of having a limited vocabulary, generally relying mostly on high frequency words to put across ideas. It is rare for a candidate in this band to give the impression that he/she is experimenting creatively with the language.

Band 2
Band 2 candidates make more lexical errors than band 1 candidates, but give the impression that they are trying to use a wider, less mundane vocabulary. It is this attempt to use a wider range of lexis which may cause more errors to be made. Typical errors include word formation (using the incorrect part of speech for its grammatical position in the utterance) and confusion regarding the use of difficult word pairs such as make/do, say/talk etc. Errors in the use of prepositions are also common.

Band 3
Band 3 candidates continue to make mistakes with prepositions, but these are less frequent than with band 2 candidates. All lexis is generally used accurately and appropriately, with only minor infrequent mistakes.

Band 3 candidates give the impression of having a wide vocabulary which allows them to communicate their views on the topic with a reasonable degree of interest and clarity.
Band 4

Band 4 candidates give the impression of having an extremely wide vocabulary which is almost always used accurately and appropriately to communicate effectively.

Band 5

Candidates in band 5 reach a standard higher than that described in band 4.

Section 12: Grammatical accuracy and the morpheme studies

12.1 Introduction

In the development of the descriptors for grammatical accuracy on the rating scale it will have been noticed that in Category 3 (tenses) and Category 9 (miscellaneous) many observations on separate aspects of grammatical accuracy were collapsed for the analysis. After the PCA these categories were further collapsed prior to the discriminant analysis. This was essential to cope with those categories for which the data was too little to deal with individually. The interview technique used to gather the data was not designed to elicit particular aspects of grammar, and it is impossible in any single piece of research to solve all data collection problems from the smallest aspect of grammatical accuracy in speech to the assessment of validity of an oral testing procedure.

However, because of the need to collapse categories, the data were analyzed a second time for comparison with the morpheme studies of de Villiers and de Villiers (1973), Dulay and Burt (1973; 1974), Bailey et. al.
(1974) and Larsen-Freeman (1975a; 1975b). The purpose of this comparison was to introduce a post-hoc test of whether the sequencing of aspects of grammatical accuracy in the rating scale band descriptors demonstrated any relationship to acquisition orders hypothesized by the morpheme studies, and to act as a second computational check on the initial interpretation of the transcript for producing the rating scale band descriptors. This check came in two stages. Firstly, five morphemes were compared with the results of the above studies using Spearman Rank Order correlations. Secondly, eight morphemes were used in order to construct an "acquisition hierarchy" which could be directly compared with the prose of the band descriptors of the accuracy rating scale. These two methods would allow a secondary, independent check upon the method of scale construction.

12.2 Step 1: rank orders

Although the data collected from twenty one interviews in individual categories was not sufficient for discriminant analysis which is a powerful multivariate tool, it was certainly sufficient for the less demanding tools of non-parametric statistics as used in the morpheme studies.

For this purpose, five morphemes were analyzed: plural (-s), progressive (-ing), past irregular, possessive and 3rd person (-s). The Group Means method (Dulay, Burt and Krashen, 1983 : 221) was used in
collecting the morpheme data. The rank orders are
presented in Table 5.18 in the Appendix to Chapter 5.
The study of these morphemes shows general agreement with
the acquisition orders which have been discovered by
other researchers. This acquisition order is presented in
Figure 5.11.

Fig. 5.11. The suggested natural order of acquisition
from the morpheme studies (Dulay, Burt and Krashen, 1982: 200)

<table>
<thead>
<tr>
<th>Group 1</th>
<th>-ing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>plural</td>
</tr>
<tr>
<td></td>
<td>copula</td>
</tr>
<tr>
<td>Group 2</td>
<td>auxiliary</td>
</tr>
<tr>
<td></td>
<td>article</td>
</tr>
<tr>
<td>Group 3</td>
<td>irregular past</td>
</tr>
<tr>
<td>Group 4</td>
<td>regular past</td>
</tr>
<tr>
<td></td>
<td>3rd person singular</td>
</tr>
<tr>
<td></td>
<td>possessive</td>
</tr>
</tbody>
</table>

It is claimed by those who support the thesis that
there is a natural order of acquisition, that sequence of
acquisition may vary within any one of these groups, but
that the groups tend to be acquired in the same sequence.
From the rank orders presented in Table 5.18 it can be
seen that the main difference between this study and the
morpheme studies occurs with the ordering of progressive
(-ing).
12.3 Step 2: An acquisition hierarchy and the accuracy rating scale

Constructing an acquisition hierarchy from observations of occurrences of grammatical features of student speech prior to the collapsing of data into larger categories for discriminant analysis should provide further evidence to suggest whether or not the process of collapsing the categories had significantly distorted the data. Once an acquisition hierarchy has been constructed, this may be compared with the band descriptors to see if there is a correlation between the sequencing of features of accuracy produced by the two methods. It must be stressed, however, that what follows describes a cross-sectional study of the use of structures from a sample of students at different ability levels. As it is not a longitudinal study, it is not capable of discovering an acquisition order. However, the study which follows does use the same methodology which has been used to date in morpheme acquisition studies, and in this context serves as a confirmation of the quantitative analysis of the transcript data presented in Sections 7 to 10 above.

Taking eight grammatical structures for which there were at least 3 obligatory cases within the data, an acquisition hierarchy similar to that in Figure 5.11 may be constructed. The eight aspects of grammar considered were: pronoun case (Category 2), Progressive (-ing) (Category 3), contracted copula (-s) (Category 9), Plural
(-s) in allomorphs /s/ and /z/ (Category 7), articles (Category 8), past irregular (Category 3), 3rd person (-s) (Category 3), and long plural /:iz/ (Category 7).

The two steps involved in producing an acquisition hierarchy similar to that in Figure 5.11 involve constructing a disconfirmation matrix and a stair matrix. In a disconfirmation matrix the first structure is listed in the row and the second in the column. The figure entered in the intersecting cell is the percentage of examples which disconfirm that the structure in the row preceded the structure in the column in order of acquisition. In the search for a natural order, fewer than 7% disconfirming cases is usually taken to mean that one structure preceded another. In the stair matrix the results of the disconfirmation matrix are ordered to show acquisition order reflected in the data.

In the Appendix to Chapter 5, Table 5.19 contains the disconfirmation matrix for the eight aspects of grammar listed above, and the stair matrix is presented in Table 5.20. In the stair matrix a "+" represents fewer than 7% of disconfirming cases, and (10) indicates fewer than 10% of disconfirming cases. The construction of the acquisition hierarchy is an interpretation of the structure of the stair matrix, and for the data in Table 5.20 it would resemble Figure 5.12. It can be seen that there does appear to be an acquisition hierarchy from this data which closely resembles the band descriptors of the accuracy rating scale.

222
Fig. 5.12. A possible acquisition hierarchy for eight aspects of grammar.

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Pronoun case</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Contracted copula (-s)</td>
</tr>
<tr>
<td></td>
<td>Long plural</td>
</tr>
<tr>
<td>Group 2</td>
<td>Irregular past</td>
</tr>
<tr>
<td>Group 3</td>
<td>Short plural</td>
</tr>
<tr>
<td>Group 4</td>
<td>Article</td>
</tr>
<tr>
<td></td>
<td>Progressive (-ing)</td>
</tr>
<tr>
<td></td>
<td>3rd person (-s)</td>
</tr>
</tbody>
</table>

Comparing this with the natural order as suggested by the morpheme studies it may be seen that irregular past and the article have been inverted, and the progressive has suffered the major change, having been relegated to the last group from the first.

It must be stressed that this analysis was carried out after the band scale descriptors had been constructed as a method of assessing the relative success of the original analysis of the data. Comparison of this order with the band descriptors yields the following results:

**Band 1:** Reference is made to all groups listed in Figure 5.12, although no reference is made to errors in the use of the contracted copula.

**Band 2:** It is stated in the descriptors that candidates will have few problems with past tense (Group 2), but will make errors in present simple and progressive (Group
4), articles (Group 4) and plurals (Group 3). However, in the descriptors no distinction is drawn between long and short plurals. The possibility of there being such a distinction was considered at too late a stage in the development of the scale for this to be investigated.

**Band 3:** Mistakes will occur "from time to time" with plurals (Group 3, but see the note above), and only errors in articles and present simple tense (Group 4) are noticeable.

**Band 4:** Only occasional article errors are referred to in the descriptor for band 4 candidates (Group 4).

It could therefore be argued that the relationship between the band descriptors and this post-hoc analysis of a limited number of aspects of grammar could be presented as in Figure 5.13 below, where a "+" is used to indicate that the band descriptor alerts raters to the possibility of students having problems with grammar which belongs to a particular acquisition group.

The comparison of five morphemes studied here and other morpheme studies shows that there is a correlation between the data used in this study and the majority of other morpheme studies, indicating that the acquisition sequence which the accuracy rating scale assumes is in keeping with findings in this field. Secondly, the acquisition hierarchy constructed from eight morphemes
Fig. 5.13. The relationship between acquisition groups identified in Fig. 5.12 and the band descriptors in the Accuracy rating scale.

<table>
<thead>
<tr>
<th>Acquisition Group</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band Descriptors</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

bears a direct relation to the sequence of band descriptors in the rating scale, thus acting as a check on the interpretation of trends observed in the data.

Section 13: Discussion

It is argued that the approach described here to the development of rating scales has resulted in a set of band descriptors in which validity has been considered prior to scale construction. Scales such as the ASLPR and ACTFL which have some degree of specificity in the descriptors may be used post-hoc to investigate validity, but do not seem to have been influenced by considerations of empirical validity from the outset.

Recent investigations into the validity of ACTFL (Dandonoli and Henning, 1990) have relied on multitrait-multimethod studies which are, of course, important. However, after training raters the results of the study may merely reflect the success of the training procedure.
That is, speaking scores may consistently correlate more highly with each other across methods than do scores on other traits, but this does not necessarily relate to the band descriptors used in the rating scales. It is quite possible that the number of bands on the scale and the band descriptors do not relate to any reality external to the scale itself.

It is not claimed that the scales developed for this study are in any way perfect, nor that improvements could not be made. What is claimed is that the empirical approach to scale development and descriptor writing provides a tool which includes the concept of validity prior to its use in a given testing situation. This in itself makes further post-hoc investigations of validity more interpretable in the light of the original data and developmental findings.
Task Design

Section 1: Factors in task design

The literature on task design concentrates on the role played by task based approaches to teaching English as a Foreign Language within the classroom. A major theme is the relationship of task to syllabus, which stresses the way in which the task relates to a statement of classroom "tactics" (Candlin, 1987: 6). The place of the task in a testing situation where one wishes to assess oral proficiency is somewhat different. Firstly there is no syllabus which will guide the production of tasks; secondly, the aim of the task is to elicit a sample of language performance which will allow raters to assess the oral proficiency of any given student taking the task rather than to create an opportunity for learning as such. The oral testing task is much more akin to Candlin's (ibid., 16) category of task for "research and experimentation" rather than the categories of learner training, information sharing or developing learner strategies. In the Testing Handbooks it is task type which most often constitutes almost the entire discussion of oral testing (Carroll and Hall, 1985; Heaton, 1975; Hughes, 1989; Madsen, 1983; Underhill, 1987; Valette, 1987). Task types are usually listed and described, occasionally with intuitive comment on the advantages or disadvantages of their use. Experimentation with task type is occasionally presented as the central focus of a
new test (Sunderland, Yixing and Barr, 1988), and task type has also been used for classifying tests (Madsen and Jones, 1981). However, there is little if any discussion of the procedures used in task design in the testing literature.

Useful criteria are provided in the teaching, rather than the testing literature for the development of tasks which could be used in an oral testing situation.

Candlin (1987: 9) suggests that a good task will be: balanced, motivating, co-operative, strategic, focused, open, structured and critical. We may assume that "structured" implies that the task has a specified format and goal, and that "open" (apparently contradictory) implies the need to allow freedom of outcome within the format and goal. Jones (1982: 84) argues that tasks designed for oral assessment should be reasonably short, well balanced, and provide equal opportunities for each participant when more than one student is taking part.

When discussing the grading of tasks, Nunan (1989: 98 - 99) suggests that the task developer take into account the complexity and length of any texts which are to be used (if any), the difficulty of the vocabulary required to complete the task, the expected speed of speech and the number of speakers, the explicitness of information, discourse structure, and the amount of non-linguistic support available. On the part of the task taker, the developer needs to be aware of the level of confidence in using the language, motivation, prior learning
experience, ability level, cultural knowledge and awareness, and linguistic knowledge. Studies assessing task difficulty are few and far between. Kenyon and Stansfield (1991) recommended field testing of tasks and the use of questionnaire data from students and raters to identify good and poor tasks. Stansfield and Kenyon (1992) attempted to scale a number of speaking tasks - or rather what look like functions within speaking tasks - described in the ACTFL Guidelines (1986). Using a Rasch Partial Credit model, they asked groups of Spanish, French and Bilingual Education teachers to assess the difficulty of a number of tasks in a study for the development of the Texas Oral Proficiency Test (TOPT). Although Stansfield and Kenyon discovered a reasonable alignment between the suggested difficulty level in the ACTFL Guidelines and the assessment of difficulty by teachers, this is still far from field testing tasks on students and assessing task difficulty from scores. It should be remembered that expert scrutiny and judgement in assessing item difficulty in more traditional pencil and paper tests is no substitute for the collection of empirical data from pre-testing and trialling. Rater 5 in this study commented that from initial analysis it would have been likely that experienced teachers and examiners would have classed Task 3 (described below) as the most difficult, but after seeing recorded interviews this assessment changed (see page 345), which is in line with Rasch estimates of the difficulty of the three tasks.
developed for this study (see page 293).

As Nunan (1989: 99) makes clear, one of the problems within the literature on task design is that the relative contribution of the factors discussed above to the success of any given task for the purposes for which it was designed is mostly unknown. He calls for further empirical research which will provide evidence to show which factors play the most important role in making a task successful.

In the absence of such information, the three tasks which were designed for data collection in this study followed the general guidelines for developing "good" tasks which are currently available from practitioners. In particular, it was felt that the tasks should be motivating for students. It is clearly important to distinguish between task type and topic, which is not often made clear in the literature. The first phase was to decide on the task type for each of the three tests, and secondly to assign a topic which would be considered interesting for the students to each of the task types.

Section 2: The three tasks

The three task types chosen for the study were:

1. Picture prompt OPI.
2. Text prompt OPI (with prior exposure to the text).
3. Group discussion.

Tasks 1 and 2 were designed to be conducted in approximately 15 minutes. Task 3 involved a 10 minute
preparation period (which was not recorded or rated), and the discussion lasted for around 15 minutes.

The topics chosen respectively for each of the task types above were:
3. Education in Cyprus.

All the material devised to conduct these tasks is contained in the Appendix to Chapter 6.

The topics were discussed with the teachers of the students who were to take the oral tests, and they confirmed that the students would probably be both interested in these topics and able to communicate on them.

In each of the tasks there were clearly distinguishable types of input. It is considered important that this is the case, as each of the oral tasks must be seen as a separate testing method for the purpose of conducting a validation study on the band scales used in assessing the students.

In the case of Task 1 the students were presented with a picture which they were invited to describe and comment upon. This was followed by the presentation of five more pictures from which they could choose one or two to comment upon. These prompts then lead into a general consideration of sport and leisure, with specific reference to the leisure activities of the students. As such, the picture prompt differed significantly from
picture tests described in Bachman (1981) and Palmer (1981) in that the picture was mainly used as a way of beginning a conversation, rather than being the prime elicitation technique.

Task 2, Population and Poverty in the Third World, involved working on the input text in class prior to taking the oral test. All students were given a passage on the subject to read in class two days prior to taking the test. The class teachers were told that the students may read the passage for approximately 30 minutes, after which the students may discuss the passage in class for a further ten minutes. All copies of the passage were then collected by the class teacher and returned to the researcher. When the students arrived to take the oral task the passage provided a starting point for discussion. Further input provided whilst the task was actually taking place included a graphic representation of the material included in the passage, and an advertisement asking for help for children in the third world.

In Task 3, the group discussion on Education in Cyprus, the students were provided with information about education in Cyprus, specifically the proposed building of the island's first university, together with a task card, note paper and pencils. The students were then given exactly ten minutes to read, make notes or prepare whatever they wished to say. At the end of the ten minute preparation time the task began.
The first two tasks were conducted by the interviewer with one student at a time. In the third task groups of students attended simultaneously, and the task involved group discussion rather than an interview as such. Within the definition of a simulation by Jones (1982) group discussion which is guided by specific input does qualify as a simulation. This makes the third task somewhat different from the other two both in terms of the number of participants involved, and in terms of their roles. In the first two tasks the interviewer played the traditional role of eliciting assessable language, whereas in the third task the role was that of "controller" rather than participant. It may be, as Jones claims (ibid., 83) that in such a task the controller is better able to assess the proficiency of the students once freed from the need to maintain interaction. It was certainly the case that the interviewer was required to do much less work in maintaining the interaction when conducting Task 3 than in conducting the other two tasks. Considerable evidence was gathered from students (Chapters 7 and 10) that they preferred Task 3, and that it provided them with a greater opportunity to speak than the other tasks, and from raters that Task 3 provided the most assessable language in terms of both quantity and quality (see page 346).
Section 3: Expected discourse outcomes

Each of the tasks was also designed in the expectation that it would generate different outcomes in terms of discourse. Task 1 on sports and leisure was designed to elicit:

* expressing personal opinions
* selecting and expressing personal preferences
* stating likes and dislikes
* narrative (past engagement in sporting events)
* description of a picture
* description of personal leisure activities

Task 2 on population and poverty in the third world was designed to elicit:

* summary and explanation of a technical text
* comparison and contrast (between Cyprus and another country)
* interpretation of graphic material
* expressing an opinion on how to solve a problem

Task 3 on education in Cyprus was a much more open-ended task than the other two tasks, and it was difficult to predict outcomes in advance. As the direction to the students becomes less guided it becomes less and less possible to predict what will happen (Wright: 1987: 49; Nunan, 1989: 48). However, it was generally expected that the task would elicit:

* expression of personal opinions
* asking for clarification
* expression of agreement and disagreement
Task design and development is an area of research in itself, but was not a major focus of this study. As such, a thorough effort to compare actual language elicited in the conduct of the tasks with the descriptions of predicted outcomes above, which would have involved the transcription and detailed analysis of interviews (as in Stansfield and Kenyon, 1992), was not done. However, when viewing a sample of tapes the language produced was informally checked against the descriptions of predicted outcomes. On this informal basis it appeared that the tasks did tap what they were expected to tap, with one exception. Little evidence was observed of Task 1 eliciting past tense narrative. The majority of students talked about present activities and avoided talking to any extent about how they began a sport or leisure activity.

Section 4: Elicitation techniques

The elicitation technique should be distinguished from task type. The latter refers to the format of the oral interview (Zappala, 1979b). The former refers to the specific prompts, be they words, pictures or other cues which are used by the interviewer to elicit a language sample (Zappala, 1979a; Lowe, 1976; 1985b).

Research which has been conducted into the format of questions, prompts and statements made by the interviewer or controller while conducting an interview or other type of oral assessment task tends to suggest that this is a
test method facet which may influence students' scores (Bachman, 1990). This aspect of task design could be termed "interviewer related error."

No research could be found which has been done on the extent to which alteration in the wording of questions or prompts in oral testing situations may affect student output and hence influence scores. However, variation in the form, speed or content of questions or prompts would fall within the definition of test method facets presented by Bachman (1990: 119).

One specialised form of oral interview has, however, been investigated in some detail. This is in the area of standardized survey interviewing when collecting information from a specified population for a given purpose. Those most commonly reported are government popularity polls. The elicitation in these interviews differs from elicitation in real interviews or oral testing tasks in that (a) the aim of the interview is to elicit a particular opinion on a given subject, and the opinion is recorded, and (b) the elicitation technique is devised in such a way as to allow the opinion to be recorded accurately.

However, it does seem likely that variations similar to these may also have an effect on the quality of data which is collected in a much more open interview or oral testing task where the aim is to elicit a sample of assessable language. Whilst oral language testing encounters are not entirely like the standardized survey
interview, they have more in common with it than, for example, conversations on television talk shows, where there is no element of assessment in an interview which nevertheless operates within certain time frames and with certain rules. Indeed, Fowler and Mangione (1990: 18 - 20) argue that in most types of exploratory research where results are to be quantified, advantages are to be gained from standardizing the elicitation technique. Any variation in the format of the interview or task which could be due to the role of the interviewer could be responsible for the variation in data which is gathered, even if this is open ended rather than the more usual closed-option format of the standardized survey interview. It seems that this would be more true for an oral testing task where the candidates are assessed on a band scale devised specifically for the task rather than band scales which could be used with any task, as the nature of the task would be severely altered. However, even with band scales which are not task specific it would still appear that great variation would constitute an alteration of the task, as has been recognised by some test developers (Calderbank and Muhammed, 1988: 53). For this reason, results from survey interviewing studies have been taken into account in the design of the tasks which were used in this study for the data collection process.

Those aspects of survey studies which relate most directly to oral testing concern the construction of
elicitation techniques and their use by the interviewer (in the case of this study, as only one interviewer was involved) or across interviewers when an oral test is administered to large numbers of candidates, perhaps in many countries.

Elicitation

Questions or prompts are sources of error when:
(a) they are misunderstood,
(b) they expect a response which the interviewee cannot give in terms of culture and/or content, and
(c) they require an answer of a personal nature which the interviewee is not prepared to give.

Interviewers

The interviewer is a source of error when s/he:
(a) alters the form or content of the question or prompt,
(b) biases the response of the interviewee by indicating his/her own views on the topic.

It would follow that in order to standardize the elicitation technique, the interviewer should stay as close as possible to the elicitation procedure as described by the designer - preferably by asking the key questions or delivering the key prompts at each phase in the task in exactly the same way for all students. After this it will be necessary to add questions or prompts which are non-standardized if the candidate takes the conversation in a particular (or unexpected) direction, but all comments or questions in between the major phases of the interview should be non-directive. Indeed, with reference to standardized survey interviews, it is
precisely in such probing activity that standardization remains almost impossible to maintain (Fowler and Mangione, 1990: 45).

In order to maintain the standardization of the elicitation procedure so that no interviewer related error appears, it is essential that the elicitation technique is described adequately and clearly, and that at key phases in the interview, the interviewer ask the same questions or use the same prompts. For this, interviewer (not necessarily rater) training is essential. Bradburn and Sudman (1979) discovered that as interviewers became more experienced the more casual they became about the way in which they altered the elicitation technique. This would also indicate the need for periodic retraining of interviewers.

It would appear that the major problem with any degree of standardization in an oral testing task would be that of encouraging the interaction to become a formal closed response encounter rather than resembling as close as possible normal conversation (Wright, 1987: 49 - 51). This could be avoided if the task designer designs the tasks in such a way that:

* the goal/purpose of the task is clearly explained to the student,
* the guidance given to the student is standardized in each administration,
* the phases in the task are started and concluded in the same way, but
* candidates are allowed to develop the topic in any way which they choose without directive/biased intervention from the interviewer/controller, and
within phases non-standardized interaction takes place in order to maintain the conversation.

The elicitation techniques used in each oral task should therefore be specified as fully as possible in terms of key prompts within phases. These are described below for each of the tasks.

Task 1: Sport and Leisure

Phase One: Warm Up
* Greet the student
* Check the name of the student
* "Tell me, X, how long have you been learning English?"
* "That's quite a long time. You must think it's important to learn English to study (+ contextual reference eg., for five years). Tell me, why do you think learning English is so important?"

Phase Two: Picture Task
* "Now, I'm sure that you've done this kind of thing before. What I'm going to do is show you a picture. You'll have a few seconds to look at it and then I want you to describe what you see and make any comments you want about it. Okay?"
* Show Picture One
* "Why do you think people take up dangerous sports like mountain climbing? (Optional: "Would you like to try this sport if you had the chance?"
* Show other pictures.
* "If you could take up one of these sports, which one would you choose, and why?"
* "Which is your favourite sport?"
* "Is any special training necessary?"
* "Generally speaking, do you think sport is important in people's lives?"

Phase Three: Wind Up
* "Do you have much leisure time when you're studying?"
* "How do you spend your leisure time?"
* Follow up any interesting points made.
* Wish students well with their studies.
* Leave taking.
Task Two: Population and Poverty in the Third World

Phase One: Warm Up

* Greet the student
* Check the name of the student
* Check that they have had time to read the passage
* "Tell me, X, have you ever visited any other countries?"
* If yes, then "How do you think life in Y differs from life in Cyprus?"
* If no, then "How would you expect life in an African or Asian country to differ from life in Cyprus?"

Phase Two: Text and Graphic Task

* Show student the previously prepared passage. The candidate may spend one minute skimming the test.
* "Could you summarise what you think is the main argument of this text for me?"
* Show student the bar graphs.
* "From these tables, what do you think are the main trends in population growth?" (Optional: "What do the death rates in these areas of the world tell us about how well off the people are?")
* If you were president of an African country such as Ethiopia, what measures would you take to reduce the birth rate?"
* Show the candidate the advertisement.
* "How is this related to the subject of the passage?" (Optional: "Do you think this kind of advertising is effective?")

Phase Three: Wind Up

* "Do you think that there is anything people in Cyprus can do to help Third World countries?"
  Develop interesting ideas briefly.
* Leave taking.

Task 3: Education in Cyprus

Phase One: Warm Up

* Greet the students
* Check names of the students
* Explain to the students that they will have ten minutes to prepare before beginning the task.
* Give each student the passage about the university of Cyprus and a Task card.
* Read the Task Card aloud to the students.
* Ask the students if they have any questions about what they are expected to do.
* Inform the students that they may make any notes they wish.

Phase Two: Discussion

* The controller should attempt to keep the length of the discussion to approximately 15 minutes.
* The controller may add a point or ask a question if necessary only to keep the discussion going, but should allow the students to maintain the interaction on their own as far as possible.

Phase Three: Wind Up

* "Would you prefer to go to a foreign university even if there were a university in Cyprus?"
* Develop any interesting points made.
* Leave taking.

Section 5: Description of the student sample

All tasks were attempted by 47 students studying in Cyprus, between January and March 1991. All students were pursuing courses leading to the First Certificate English Examination of the University of Cambridge Local Examinations Syndicate.

The average age of all students was 15 years, five months. The number of female students was 21, and their average age was 15 years, 7 months. The total number of male candidates was 26, and their average age 15 years, 2 months.

The students were identified as coming from three subgroups of the population in advance of the study taking place, based upon their class placements and teachers reports of their language learning history. Group 1 was labelled "Good language learners", Group 2
was labelled "Average language learners" and Group 3 "poor language learners". Their numbers and average ages are given below.

**Group 1**

Total number of students: 11  
Average age: 15 years, 0 months  
Number of female students: 6  
Average age: 15 years, 0 months  
Number of male students: 5  
Average age: 15 years, 1 month  

**Group 2**

Total number of students: 17  
Average age: 15 years, 3 months  
Number of female students: 9  
Average age: 15 years, 3 months  
Number of male students: 8  
Average age: 15 years, 2 months  

**Group 3**

Total number of students: 19  
Average age: 16 years, 6 months  
Number of female students: 6  
Average age: 16 years, 7 months  
Number of male students: 13  
Average age: 16 years, 5 months  

All students in the sample were native speakers of Greek, although one female student from Group 1 was bilingual in Greek and Hungarian.
Section 6: **Conduct of the oral tasks**

All the oral tests were conducted during normal classroom time, each student in turn leaving the regular classroom and going to a room set aside for the purpose of taking the oral tests. The students were not allowed to talk to each other about the oral tests in the classroom, and were supervised by a member of the administrative staff whilst waiting to take each of the tests. Groups of students who did know each other took each of the tests in the same order and on the same day so that communication between them concerning the test was virtually impossible. The room was arranged with comfortable chairs set out around a coffee table, the video camera was positioned at the far side of the room, pre-focused on the area in which the tests were to be conducted so that an operator would not be required. The room in which the oral tasks were conducted and the interviewer were the same on each occasion, holding constant at least these test method facets in order to reduce the number of potentially confounding factors in the validity study.

Each test was recorded on video tape for future assessment. One third of the candidates began with Task 1, one third with Task 2 and One third with Task 3, although the beginning and completion of the 3 task sequence was necessarily staggered, as only one interviewer was being used. This design rules out the possibility of order effect.
Immediately upon completion of all three tests the students were given a detailed questionnaire to complete regarding their attitude towards each of the tests and asking them to assess their own level of oral proficiency. Eight of the forty seven students were shown video recordings of one of their tests two weeks after the completion of the study, and were asked to retrospect regarding their performance. These interviews were recorded on audio tape.

Each student took all three tasks within the space of one and a half weeks, and it took six weeks in February and March 1991 to conduct all tasks for all students.

Two problem cases arose. One student suffered from such severe anxiety that she could only whisper during the tests, and raters reported that they could not hear her clearly enough to award scores from the video tapes. This student was therefore excluded from further analyses, reducing the sample to 46. One student was taken ill immediately after the completion of the three tests, and did not complete a questionnaire. This meant that the analysis of questionnaire data is from a total of 45. This did not, however, influence the concurrent study as FCE scores were available for the student.

The picture quality of the video data was adequate, although not good enough for the video cassettes to be used for rater training, for example. However, as rater training was not an issue in this study the picture
quality is considered to be unimportant. The audio reproduction was good. Only one student could not be assessed from the recorded data, as reported above.
Chapter Seven

Student Affective Factors

Section 1: Introduction

In this study we are interested in student affective factors because of their potential to confound results in the reliability and validity studies which are described in Chapters 8 and 9. If affective responses to test situations or tasks are seen to account for a significant proportion of variance the results of reliability and validity studies are called into question, as even positive results could be caused by systematic affective factors rather than student ability as recorded on the rating scale being used. This Chapter is therefore concerned with establishing that affective factors did not influence results to any significant degree, and that these factors can be eliminated from consideration as confounding factors in the evaluation of reliability and validity.

Testers often claim an interest in the responses and reactions of students to testing situations. Underhill (1987) makes this one of the central themes in his book on oral testing much applauded by Lantolf (1988), and yet there is no reference at all to research on factors that might play a role (Fulcher, 1990: 81). Other researchers, notably Ingram (1985b) specifically isolate factors such as introversion / extroversion, intelligence, experience, education, social norms and willingness to communicate, as factors that may influence student affective reactions
to the oral testing situation. The underlying problems raised by work dealing with the nature of communication in the oral test itself (van Lier, 1989; Perrett, 1990) and choice of topic (Douglas and Selinker, 1985) are clearly also related to affective aspects of testing, in that these factors may systematically influence the behaviour of the student under test conditions independently of the oral trait that the tester wishes to measure. The extent to which this is reflected in test scores represents a confounding factor in the interpretation of those scores. It is therefore the responsibility of the test researcher to ensure that some of the most common confounding factors have been accounted for prior to interpretation of results.

Shohamy, Reves and Bejarano (1986, 213) state that in developing a test of oral proficiency in Israel, one of their aims was to "evoke positive attitudes on the part of the test takers." This aim is important in that such a positive attitude would, it is suggested, reduce the number of potentially confounding factors that might affect oral scores. Shohamy and Stansfield (1990, 87) report that the project was successful in this respect, with subjects indicating a positive attitude towards "content, technical quality and ability of the HeST [Hebrew Speaking Test] to probe their speaking ability..." This is in line with similar results (Shohamy, 1983a, 235 - 236) which suggested that direct speaking tests were considered to be accurate measures of
ability by students themselves. They enjoyed them, were generally comfortable, and found the experience interesting and challenging.

The other major issue which has received considerable attention is that of test anxiety. Galassi et al (1984) and Madsen and Murray (1984) have shown that a small but significant amount of score variance can be attributed to test anxiety. Savignon (1972, 1985), Ingram (1985b), and Shohamy and Stansfield (1990, 87) suggest that anxiety is reduced in a direct testing situation such as an oral interview because of the presence of the human interlocutor, as opposed to a tape recorder in a semi-direct test of speaking.

Scott (1986) investigated affective factors in relationship to a pair oral interview and a group oral interview, and concluded that there was no essential difference between questionnaire responses of students who had taken the one test format or the other. However, similar concerns were expressed by both groups especially with regard to time constraints.

Young (1986) suggests that subjects of low ability suffer more from debilitating anxiety than high ability students, but discovered a small significant negative correlation between assessments of anxiety and ACTFL scores, suggesting that some anxiety does facilitate performance in an oral interview. Young's conclusions (ibid., 439) are in line with those of Galassi et al (op. cit) and Madsen and Murray (op. cit) that "ability, not
anxiety, is the more important variable affecting OPI scores. Similar conclusions are reached by Barley (1990), who found non-significant correlations between oral test scores and reported as well as measured anxiety levels. However, the research done on the relationship between anxiety and test scores does not distinguish between trait, state and situational anxiety as is generally the case with anxiety studies in other areas (MacIntyre and Gardner, 1991).

Section 2: Questionnaire structure

As soon as the students used in this study had completed all the oral tasks, they were asked to complete a questionnaire for each of the tasks, with a few general questions attached to the end. The questionnaire for Task 1 is included in Appendix 7.18, together with the general questions. This was adapted from that used in Scott (1986). There was no difference between the questionnaire format used for each task.

In order to investigate the structure of the questionnaire, Principal Axis Factor Analysis was used. In the first analysis only factors with eigenvalues of 1 or greater were extracted, and this produced 2 for each of the three questionnaires, accounting for 24.95%, 33.17% and 40.12% of variance respectively. Initial eigenvalues indicated, however, that another two factors could be extracted with eigenvalues of greater than .5, and so this option was taken in order to extract the
maximum amount of information from the data available (Farhady, 1983: 18 - 19).

Tables 7.1, 7.2 and 7.3 in the Appendix to Chapter 7 contain the results of the Factor Analysis for the three questionnaires with 4 factors extracted. In the three analyses there are many consistencies. The first is that questions 2 and 3 always load on the same factor, and it is these two questions that are concerned with test anxiety. On questionnaire 1 and 3, question 5 also loads on the same factor, and this question also seems to relate to anxiety in that it is concerned with whether or not the students believe they would have done better had they taken the test on another day. It would seem that questions 2, 3 and 5 make up a block concerned with test anxiety.

The second consistency is that questions 1 and 6 load on the same factor on questionnaires 2 and 3, although in questionnaire 1 question 6 stands alone on factor 4 and there are no significant factor loadings for question 1. Both of these questions concern the students’ perception of the validity of the task as a test. The fact that these questions cluster is fully in line with Scott’s (1986) view that students are capable of separating perceptions of the validity of a test from their perceptions of test difficulty.

A third cluster of questions is based around questions 8, 13 and 14, all of which load on the same factor in questionnaire 1, and co-occur with questions 9
and 15 in questionnaire 2 and 15, 4 and 6 in questionnaire 3. Question 8 concerns whether or not the student liked doing the task, question 13 concerns whether or not the student thinks that the task was interesting, and question 14 with whether or not taking a task was an unpleasant experience. This cluster therefore seems to be an "enjoyment" factor. It is noteworthy that the only other question that loads on the same factor in two of the questionnaires, question 15, is concerned with the students' perception of the fairness of the topic chosen for the interview. It would therefore seem that the choice of topic is related to perceptions of enjoyment of the experience of taking the test, rather than perceptions of validity or indeed test anxiety.

From the factor analysis, it is believed that there are three coherent factors at work in the student responses to the questionnaire. These are the factors of test anxiety (questions 2, 3 and 5), perception of test validity (questions 1 and 6) and enjoyment of the test experience (questions 8, 13, 14 and 15). When looking at Tables 7.1 - 7.3 it should be remembered that negative loadings frequently reflect reverse scoring of the question. Apart from question 14 these are not high, and do not appear to form any interpretable patterns.

Section 3: **Summary statistics for the questionnaires**

The mean and standard deviation for each question for 46 respondents are presented in Table 7.4 in the
Appendix to Chapter 7. Prior to commenting on question groups, it is interesting to note that the responses to question 11 consistently show that students did not on the whole think that the interviewer was a factor in the scores that they receive. However, it is on Task 3, the group discussion, that students think the interviewer is less important, that they have more control over their own destiny, as it were.

It is now possible to look at the students' responses within the factor blocks for each task, beginning with the part played by anxiety.

Student responses indicate that there was a fair degree of anxiety prior to doing task 1, although not whilst doing the task. With task 2, there would appear to be much less anxiety before taking the test, whilst the group discussion appears to produce the least amount of anxiety. This result is not an order effect, as the order was rotated.

When it comes to validity, although there is agreement amongst students that task 1 provided them with an opportunity to demonstrate their ability to speak English, in general students were much less positive about their perception of the validity of task 1 in providing the examiner with accurate evidence of their ability. This must be compared with task 2, which was the most difficult task for all students (see page 293), which the students see as providing accurate information for the examiner. Again, task 3 is viewed positively by
most students as a valid test. The first task, the picture prompt, was not seen by the students as a valid test of their oral ability.

Task 1 was generally considered to be an enjoyable experience, as was task 3. However, more students thought that task 2 was not enjoyable and that the topic was difficult. Task 2 is a good example of an oral testing task which many students did not enjoy, but view as a valid test of their oral ability. This illustrates the way in which students are able to make distinctions in their value judgements of what they are expected to do in examination conditions.

Section 4: Analysis of questionnaire responses by group and task preference.

One way ANOVAs were conducted for each of the three ability groups discussed on pages 242 - 243. Where the ANOVA was significant at p < .05 a post-hoc Tukey test was conducted to discover the source of the difference in response patterns.

For task 1, differences were detected in questions 5 and 6. For task 2, differences were detected in questions 3 and 13. For task 3, no differences in response patterns were noticed.

Question 5 related to anxiety, and question 6 to perceived validity, question 3 to anxiety and 13 to enjoyment of the task. The results are presented in Tables 7.5 to 7.8 in the Appendix to Chapter 7.
Groups 2 and 3 tend towards having no opinion regarding whether or not they would have done better on task 1 had they done it on another day.

In this case, the less able students considered task 1 to be capable of providing evidence upon which the examiner could accurately assess their oral abilities. As proficiency increases, the perceived validity of the task decreases. Group 1 students (more able) actually disagree with the view that task 1 is capable of eliciting an oral sample upon which they can be fairly assessed. It will be shown later (see page 293) that task 1 was the second most difficult task, after task 2. However, the perception of more able students was that it was too easy for them.

Less able students from Group 3 report suffering from anxiety whilst doing task 2, whereas the other two groups did not tend to suffer from anxiety. Again, it will be shown that task 2 was by far the hardest on all rating scales, and it would appear that for less able students this was a stressful experience. From these results the importance of selecting a task that has a difficulty value not very much greater than the average ability level of the students would seem to be an important factor in reducing test related anxiety. In this respect, Rasch partial credit models do seem to offer a great advantage over other methods of test analysis, for as Rasch models measure task difficulty and student ability on a common scale, it is possible to
select a task the difficulty of which is estimated to be at a level which would not induce stress in the students. This would reduce test anxiety to a minimum and remove a potentially confounding factor from the testing process.

Even though less able students reported being anxious when they were taking task 2, Table 7.8 shows that they did find the task interesting. As it was discovered that question 13 was related to the concept of enjoyment of the task together with questions 8 and 15, summary statistics are also presented for these questions by group in Table 7.9 in the Appendix to Chapter 7.

Although there is some variation in the responses of Groups 1 and 2, it is quite clear that Group 3 students consistently report enjoyment of task 2, despite anxiety and difficulty.

It can only be concluded from these results that the concepts of anxiety and enjoyment are separate, and that the students are indeed capable of distinguishing between them. A reasonable degree of anxiety does not appear to detract from any enjoyment less able students are capable of experiencing from the oral test.

Question 16 on the questionnaire allowed each student to select the task that they would prefer to do if they were given an option within an oral examination. The number selecting each task is given in Table 7.10 in the Appendix to Chapter 7.

Easily the most popular task was the group discussion, task 3, with slightly over 50% of students
stating that they would opt for this test format if given a choice. Task 1, the picture prompt was second, and task 2 the third option for all groups of students.

As the students can be classified into three ability groups and three task preference groups, it is possible to use multivariate analysis of variance (MANOVA) on the scores for Fluency, Accuracy and the ELTS score for each of these classifications, in order to investigate whether a preference for a particular task or an interaction between ability level and preference account for the scores received. In the Appendix to Chapter 7, Table 7.11 contains the MANOVA results for the ability group variable, Table 7.12 for the task preference variable, and Table 7.13 for the interaction.

Table 7.11 confirms the results of the Group Discrimination validity study on page 315, although it should be noted that on task 3 the Accuracy scale does fail to discriminate amongst students. This is probably nothing to do with the task, but with the scale itself, as it will be shown later that it is used inconsistently across tasks.

In the case of the Task preferences, none of the univariate test statistics are significant, although the multivariate statistics indicate that prediction of preference is possible when using all sources of information available. This suggests that there may be some interaction effect between group and preference.

From Table 7.13 it can be seen that there is an
interaction effect of Group and Task preference for the Fluency scores on tasks 2 and 3, for the Accuracy score on task 2, and for the ELTS score on task 2 and 3. To be able to spot these interaction effects, means and standard deviations from all students classified by Group and Task preference are given in Tables 7.14 and 7.15 in the Appendix to Chapter 7. It should be noted that further statistical analysis of this data cannot be carried out, as the total number of 45 respondents has been broken down into categories, some of which are too small to provide reliable results. Some categories also have no variance, making interpretation especially difficult. However, the following possibilities do suggest themselves. (a) Students from Group 2 who preferred task 3 tended to get higher scores on all scales on task 2, and higher than usual scores on the Fluency scale on task 3. (b) Students from Group 1 tended to score higher on the Accuracy scale for task 2 when they state they prefer task 2, and lower than usual on the Fluency scale when they state that they prefer task 3. (c) Group 3 students score less than usual on the Accuracy scale on task 2 if they prefer task 3, but better than usual on the ELTS scale for task 2 if they state they prefer task 2.

Two of these observations (b) and (c) are consistent with the possibility that student preferences for test format at certain ability levels do affect the scores that they will get to some limited extent. Observation
(a), however, contradicts this, suggesting at least for middle ability level students their preferences for test format are unlikely to have any influence on scores at all.

Section 5: **Student self assessment**

Question 18 on the questionnaire asked the students to assess themselves in general oral proficiency, grammatical accuracy and fluency, on a five point scale ranging from very good to very poor. One way ANOVAs were performed to test whether there was any difference between the responses of the three groups to the question, and in each case these were not significant, suggesting that students are incapable of assessing their own oral proficiency. This is a similar finding to that of Bohn and Bohn (1985) who found that students were not capable of accurately assessing speech samples of their own or their peers within a classroom setting.

The responses to the three part question were correlated with the Rasch estimates of ability from the Fluency scale, Accuracy scale and ELTS scale, partialling out task effect. The results are presented in Table 7.16 in the Appendix to Chapter 7, and confirm the view that students do not seem able to assess their own oral proficiency.
Section 6: The relationship of reported anxiety to Rasch ability estimates

The findings of this research are not completely in line with those of previous studies concerning the relationship of anxiety and oral test results, in that there is a weak correlation between lack of anxiety and better test results (Scott, 1986: 107). The data are presented in Table 7.17 in the Appendix to Chapter 7. It should be noted that although the correlations are inverse, questions 2, 3 and 5 were scored with "strongly agree" = 5. The correlations between reports of anxiety and slightly higher test scores should therefore be read as positive.

The results of this study, however, do not contradict the view that a limited amount of anxiety facilitates student performance in oral tests, for the correlations are too weak to draw any firm conclusions.

Section 7: Summary of qualitative analysis of responses to open ended questions, for questions 6, 8, 12, 15 and 17

Question 6 on each questionnaire was concerned with the perceived validity of the task, and question 8 with enjoyment. Question 12 was intended to tap perceived difficulty of the task, but this did not relate to any other questions in the factor analysis. Question 15 also concerned perceived enjoyment of doing the task, whilst question 17 attempted to discover what, if anything, would make the oral testing situation less stressful for
the student.

Student responses to these questions will not be analyzed in detail, as they do not relate specifically to the assessment of rating scales. However, interesting responses may be summarised in the following four categories.

**Category 1: Interviewer related responses**

Most of these concerned the person interviewing the students or conducting the oral test, and fell roughly into two categories: comments concerning the interviewer him/herself, and comments concerning the examiner and the organisation of the oral test. In the first category came comments such as interviewers should provide students with more help in order for the latter to continue talking if they feel that they cannot say anything else, that the interviewer should make more jokes to help students to relax, the interviewer should phrase all questions clearly, and finally, the interviewer should be a "kind" person who allows and encourages the students to speak openly and freely. In the second category came comments such as there should be time prior to the oral test for the students to meet the interviewer "socially" – one student even said that the interviewer should offer drinks to students prior to the oral test beginning – so that they could get to know him/her better before the test thus reducing stress associated with having to talk with an unknown person. Two students would
have been happy with a five minute "chat" prior to the interview.

It would appear that part of the stress associated with an oral test does seem to be the foreboding connected with having to talk for a given length of time to a person whom one has never met before.

**Category 2: Technology related responses**

Three students reported feeling nervous during the interviews simply because there was a video camera in the room. This point is to be taken seriously, as recording equipment may make some students more nervous than they would otherwise be. However, this cannot be helped when it is necessary to grade after the interview or when recording is used for moderation or research purposes.

**Category 3: Task related responses**

Two students strongly believed that students should be provided with information on the format and subject of the oral test prior to the test taking place. Not knowing the subject or precisely what would be expected of them was the prime cause of anxiety for this small percentage of students.

**Category 4: Student related responses**

Four students specifically commented that anxiety had nothing to do with the testing situation or the interviewer, but with their own state of mind.
"Uncertainty in myself" was one reason for nervousness, and another student said that the only way to feel less anxious was to study harder, that "knowing more English and being more fluent" was the only way to reduce anxiety. Two students said that nothing could be done to reduce anxiety; anxiety for them was a natural part of any test, and that they would always be anxious immediately before the start of the test, but that they consciously try to relax and usually do this after the first few minutes.

Section 8: Discussion

From the answers to the questionnaire, it may be claimed that these students are capable of distinguishing clearly between their enjoyment of doing an oral task, their view of the validity of the task as a method of assessing their language abilities, and their anxiety in the testing situation. They are also capable of expressing their views about what factors are important in improving the tasks. Task one was generally seen to be fair, enjoyable, but invalid as a test for all but the weakest candidates. Task 2 was seen to be valid, but not enjoyable, causing a fair degree of anxiety. Task 3 was seen as the fairest test, enjoyable and valid by most students.

The relationship between anxiety, perceived validity, task difficulty, task preference, task enjoyment and student ability is complex. Task 3, the
Group Discussion appears to be the best test of oral abilities from the point of view of utilising positive student affective factors, as it is viewed as the most enjoyable, the least difficult, the most valid and creating the least amount of test anxiety. It is therefore not surprising that most students said that they would prefer to take task 3 in an oral test if they were given a choice. The situation with regard to task 1 (picture prompt) is not quite so clear. Although it was the second preference of students, and considered to be enjoyable, it was viewed as not being a valid test of the speaking ability of more able students, and causing considerable anxiety to all students. Task 2, on the other hand, was the least preferred of the three tasks, and was considered not to be enjoyable although it was viewed as a valid test of oral abilities by students generally.

Task preference does seem to be inversely related to task difficulty, but there would appear to be little obvious connection between other factors at work. However, the two exceptions to the general picture presented above, namely the variability in results from lower ability students in seeing task 1 as a valid test and registering high anxiety but enjoyment while doing task 2, may indicate that although lower ability students generally enjoy oral tasks irrespective of difficulty and anxiety, they prefer tasks which they perceive as being easier.
That there does not seem to be any strong relationship between these factors is not surprising, as none of them have a great deal of effect upon test scores. As Table 7.12 in the Appendix to Chapter 7 shows, the most important factor affecting test scores is that of student ability. We may therefore conclude that the results of the following reliability and validity studies, which are based upon test scores, are not confounded with affective factors which would render their interpretation impossible.
Chapter Eight

Reliability

Section 1: Introduction

In this chapter the issue of the reliability of the Accuracy scale, the Fluency scale and the ELTS global rating scale will be addressed. To establish the reliability of the two rating scales developed in this study is particularly important, for unless scores can be seen to be reliable the investigation of validity becomes impossible. That is, if the amount of error variance associated with a rating scale is high the researcher cannot have confidence that the rating scale is measuring anything, and so it makes little sense to question the nature of what is being investigated. In a reliability study it is therefore important to estimate and account for any variance which is not relevant to the ability of the students in the scores awarded on the rating scale.

The reliability of the testing process developed was analyzed using three separate methods. Firstly, a G-Study was carried out following the methods outlined in Bolus, Hinofotis and Bailey (1981), Cronbach (1984: 161 - 163), Crocker and Algina (1986: 157 - 185), Feldt and Brennan (1989: 128 - 136) and Bachman (1990: 187 - 197). Secondly, traditional inter- and intra- rater reliability was estimated using correlations. Thirdly, a Rasch Partial Credit model was used to investigate variations in rater behaviour and rater severity.

The data were provided by five raters who rated all
47 students on the three tasks on three rating scales. The five raters were all trained teachers of English as a Foreign Language working in Cyprus during the academic year 1991 - 1992 (first degree and RSA Diploma in the teaching of English as a Foreign Language), and were selected for this study on the basis of their interest in oral testing. Each of the raters is referred to by a number from 1 to 5 in what follows as their anonymity was promised. Raters 3, 4 and 5 were trained, practising IELTS examiners, whilst raters 1 and 2 had no previous experience of oral testing.

The raters received no training in the use of the rating scales at all. Each rater was provided with a copy of each rating scale, a marksheet on which to record ratings, and an information leaflet (one single sheet of A4 paper) on which they were asked to "study the rating scales before using them" and which stated that they should rate each student taking each task simultaneously on all three rating scales from one viewing. Given the length of the video tapes it was not considered feasible to ask volunteer raters to spend more than 40 hours on the task of rating. Finally, the information leaflet asked the raters to keep notes on the tasks, the rating scales, and any students with whom they encountered problems in the rating process.

The video tapes were given to the raters at the beginning of August 1991 and returned by mid-September 1991. All the raters were allowed to complete the rating
task in their own homes during this six week period.

Section 2: Generalizability study

The advantage of conducting a G-study to assess reliability is that it is capable of taking into account facets of the testing situation and the extent to which these may affect the scores which students are awarded. In this study Fluency, Accuracy and ELTS global ratings of ability were treated as conditions, whilst rater and task were treated as nested facets, the former consisting of five individuals and the latter consisting of three tasks. Variance components can be assigned to each of these facets and conditions, and reliability of the testing process calculated as the extent to which results could be generalized beyond the instances of each specific facet under which the scores were obtained. In order to obtain these variance components ANOVA is used. As the test method facets are specified in the ANOVA, it is the error or residual component which is interpreted as person variance and hence reliable variance. The assumption being made is that no facets other than those specified in the ANOVA affect test scores, which is justified to the extent that affective factors do not appear to confound results, as demonstrated in Chapter 7.

This design generates three ANOVA tables, one for each of Fluency, Accuracy and ELTS global ratings. Each of these was followed by a post-hoc test of the effect of specific components of facets where unreliability was
Fluency

Table 8.1 in the Appendix to Chapter 8 indicates that the rater is the largest source of unreliability in the assessment of fluency, but this variation is not large enough to reduce the inter-rater generalizability coefficient to unacceptable levels. There is also significant influence from task, but as a percentage of total variation within the assessment of fluency this is also fairly small. However, these differences were considered to be large enough to warrant further investigation.

A post-hoc test was carried out on the raters and the tasks in order to ascertain the precise cause of the unreliability. In each case it was hypothesized that there was no significant difference between the individual variances assigned to raters and tasks.

Table 8.2 in the Appendix to Chapter 8 presents the post-hoc analysis of difference between raters, and a pattern of rater agreement emerges from the data which may be summarised as in Figure 8.1, where a "+" indicates no significant disagreement and a "-" indicates significant disagreement.

The greatest source of unreliability when assessing fluency is seen to be rater 3, whose rating patterns are analyzed more carefully later. Rater 5 also differs significantly from other raters.
The influence of the tasks was analyzed in the same way as the raters. Table 8.3 in the Appendix to Chapter 8 shows that although the unreliability introduced by task was small, this is accounted for almost entirely by Task 3 which does seem to be operating in a very different way from the other two. This is shown in Figure 8.2.

The nature of Task 3, the Group discussion, will be discussed later in the light of all investigations into the operation of the task test method facet.

Accuracy

A very similar pattern of results may be observed when analyzing the reliability of the use of the Accuracy scale. However, the test method facet variance components
shown in Table 8.4 in the Appendix to Chapter 8 are slightly higher than in the analysis of the Fluency scale, and the reliability coefficients thus reduced.

The post-hoc analysis in Table 8.5 in the Appendix to Chapter 8 shows a data pattern which is almost identical to that produced from the analysis of the Fluency scale. This indicates that there is a large degree of consistency in the extent to which raters converge or diverge in rating patterns across scales. This is not an argument for inter-rater reliability, but for intra-rater consistency which, given that the analysis is across scales, may indicate a degree of cross-contamination in the rating process. Rater 3 remains the largest source of unreliability in the assessment of accuracy, as can be seen from Figure 8.3.

The pattern observed in the analysis of the Fluency scale is also repeated for variation between tasks, with Task 3 showing that it operates somewhat differently from the other two, as shown in Table 8.6 in the Appendix to Chapter 8 and Figure 8.4.

Fig. 8.3. Rater agreement (Accuracy)

<table>
<thead>
<tr>
<th>Rater</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>
Table 8.7 in the Appendix to Chapter 8 shows that the inter-rater coefficient is higher for global assessment than for the other scales, but when the post-hoc tests were carried out the amount of unreliable variation between raters was no better than in the use of the other two scales. However, the pattern of convergence and divergence between raters is greatly changed, as can be seen in Table 8.8 in the Appendix to Chapter 8 and Figure 8.5.

This particular pattern is all the more puzzling because of the loss of agreement between raters four and five. Three of the raters used in this study are trained and experienced ELTS examiners: raters 3, 4 and 5. Rater
3 now shows no significant difference with rater 2, who is not a trained ELTS examiner, whilst raters 4 and 5 now appear to rate in significantly different ways.

The post-hoc tests in this case have shown that the global inter-rater generalizability coefficient is not sufficiently powerful to be quoted alone as a reliability coefficient, for it may mask a large amount of variability which still exists. This variability may be further analyzed through Rasch techniques.

Analysis of tasks as a significant test method facet revealed precisely the same pattern as previously observed with both the Fluency and Accuracy scale, as evidenced by Table 8.9 in the Appendix to Chapter 8 and Figure 8.6.

Fig. 8.6. Variation between tasks (ELTS Global Assessment)

<table>
<thead>
<tr>
<th>Task</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Section 3: Correlational analysis

Despite problems with the traditional use of correlations as reliability coefficients (Krzanowski and Woods, 1984) they are reported here as they confirm the observations from the G-study. Information which is not provided by correlation coefficients (means and standard deviations) is taken into account elsewhere.
3.1 Reliability coefficients on scales

Firstly, descriptive statistics and correlation coefficients were produced on each scale, ignoring the significant task effect discovered in the G-study (scores on each scale were averaged across tasks). A different matrix taking into account task differences is analyzed in section 3.2.

Fluency

The correlations presented in Tables 8.11 and 8.12 in the Appendix to Chapter 8 show a reasonable amount of agreement among the raters, which the G-study has shown does not represent a complete picture. The correlations ranging from .69 to .71 for rater 3 are not infrequently taken to represent a reasonable degree of inter-rater reliability, but this hides the greater leniency of this rater as shown in Table 8.10 in the Appendix to Chapter 8, and the apparent lack of ability to discriminate between higher and lower ability students.

The average correlation coefficient for all raters is .83. To obtain a reliability coefficient this figure was corrected for the number of raters using the Spearman-Brown Prophecy Formula (Henning 1987: 82 - 83), where the Fisher Z transformation is used to convert the Pearson Product Moment correlation to an interval scale, which can then be averaged and transferred back into a correlation coefficient, and the result put into the reliability formula:
\[
\text{rtt} = \frac{nrA,B}{1 + (n-1)rA,B}
\]

where rtt = inter-rater reliability, n = the number of raters, and rA,B = the average correlation coefficient. The resulting reliability coefficient is .96. This demonstrates that the concern shown in the oral testing literature for using more than a single rater (page 44) is justified, although reporting only reliability coefficients based on correlations corrected for the number of raters is clearly unjustified, as they hide even further differences among raters. However, these corrected correlations are presented here as confirmation by more traditional methods of the high reliability coefficients achieved in the G-study using five raters. This helps to remove any hesitation in using pooled scores in the investigation of the validity of the rating scales in Chapter 9.

Accuracy

Tables 8.13 to 8.15 in the Appendix to Chapter 8 present the same information for the use of the Accuracy scale. Rater 3 stands out as being more lenient and less discriminating than all other raters, and there appears to be less agreement in the rating of Accuracy than Fluency. The average correlation coefficient for all raters is .79, but when this is corrected for the number of raters rtt = .95.
ELTS Global Assessment

The analysis of the ELTS scale as shown in Tables 8.16 to 8.18 in the Appendix to Chapter 8 reveals a lower average correlation of .76, rtt = .94. This indicates that the amount of variation isolated in the G-study (see pages 268 - 273) requires further analysis using more sensitive tools.

3.2 Reliability coefficients on scales across tasks

As the G-study discovered a task effect, it was decided to look at the correlation coefficients between raters by task. This will throw light on the nature of the task effect itself. The Pearson Product Moment correlations for Fluency on Tasks 1, 2 and 3 are in Tables 8.19, 8.20 and 8.21 respectively. For Accuracy the results are in Tables 8.22, 8.23 and 8.24, and for ELTS Global ratings in Tables 8.25, 8.26 and 8.27 respectively for Tasks 1, 2 and 3. All these tables are contained in the Appendix to Chapter 8.

For the Fluency scale, the mean correlation for Task 1 was .70, for Task 2 a lower .58 and for Task 3 .72. Corrected for the number of raters these figures provide reliability coefficients of .92, .87 and .93 respectively. If this pattern were to be repeated for all scales, given that from the G-study Task 3 caused the task effect which was isolated to be a source of unreliability, it may be necessary to reassess the precise nature of task 3 and the effect which it has on
scores.

For the Accuracy scale the mean correlation for Task 1 was .63, .66 for Task 2 and .69 for Task 3. When corrected for the number of raters, the reliability coefficients are .89, .91 and .92 respectively.

On the ELTS Global scale, average correlation coefficients were .62 for Task 1, .57 for Task 2 and .70 for Task 3. These provide reliability figures of .89, .87 and .92 respectively when corrected for the number of raters.

This evidence is contradictory to that provided in the G-study. This would tend to suggest that raters are more consistent in their rating on any scale when judging student performance on Task 3, and least consistent when judging performance on Task 2. In order to investigate this aspect of reliability further it is necessary to turn to the more sensitive tool of the Rasch Partial Credit model.

Section 4: Rasch Partial Credit analysis

4.1 Introduction

The use of Rasch Partial Credit techniques in the analysis of data stemming from tests which involve judgements made on a scale is growing rapidly (Adams, Griffin and Martin, 1987; Pollitt and Hutchinson, 1987; Masters, 1990; McNamara, 1990, 1991; Pollitt, 1991). The basic model for Rasch analysis is to be found in Wright and Masters (1982), and the software used in this study
is by Linacre (1991), and the multi-faceted model is described in Linacre and Wright (1990).

Rasch techniques allow much more detailed analysis of data from rating scales, making it possible to ask questions concerning individual raters (for example, the precise nature of the assessments of Rater 3 whom we have already noted to be unreliable), individual candidates (if we should feel that they have been unfairly graded), and individual scales or tasks to investigate whether or not the band thresholds are consistent or variable across facets of the measurement situation.

4.2 Rating scales and rater bias

It should be noted that Rasch techniques do not use extremes scores on any scale in the calculation of measurement properties of a scale, and as such the order of harshness/leniency of some judges may be slightly different from those in traditional descriptive statistics. In this data, there are no changes at all. Tables 8.28, 8.29 and 8.30 in the Appendix to Chapter 8 are extended versions of Tables 8.10, 8.13 and 8.16 including Rasch estimates of judge severity. The lower the logit value the more lenient the rater and, conversely, the higher the logit value the harsher the rater.

It must be stressed that these descriptive statistics do not take into account the significant task variable, but it can still be seen that rater 3 is
undoubtedly the most lenient in the use of all rating scales. The pattern of severity in the use of the scales is wholly consistent, except for the use of the ELTS global scale, where the severity order of the judges changes markedly. This can be seen in Figure 8.7 (where Fl = Fluency, Ac = accuracy and El = the ELTS global scale).

Fig. 8.7 Rank order of judge severity on all rating scales

<table>
<thead>
<tr>
<th>Rater 1 Logit</th>
<th>Rater 2 Logit</th>
<th>Rater 3 Logit</th>
<th>Rater 4 Logit</th>
<th>Rater 5 Logit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fl. 2</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Ac. 1</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>El. 3</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Before looking at specific reliability problems, the expected scores of any student on any given scale can be mapped. At present this is without reference to any specific rater or any specific task. That is, we are looking at the probability of achieving a given score on each rating scale irrespective of any problems with reliability. The range of potential scores are presented in Figure 8.8 on the following page. The first and last line are the ability levels in logits, and the scores expected on the scales at any ability level are presented in the order: Fluency, Accuracy and ELTS global assessment.
It is now possible to look in greater detail at specific aspects of unreliability in terms of the extent to which individual raters significantly deviated from expected scores in specific cases. Misfitting ratings are presented in Table 8.31 in the Appendix to Chapter 8, where the category is the actual band score which the rater awarded the student, whose number is listed in the last column of the table. The column headed "step" is the point at which a student of a given ability moves from being in one category to another; for example, on Fluency, step 4 is the point at which a student would move from receiving a band 4 to a band 5. The column headed "Exp." is the expected score in terms of the steps, and "Resid." the difference between the expected position of the student on the steps and the actual position on the steps.

It is immediately obvious that all misfitting ratings come from rater 3. These misfitting ratings (after leniency is taken into account), with the exception of person 32 on the ELTS scale, represent cases
where the student was of higher expected ability, and rater 3 awarded a lower band. Hence, all figures in the Residual column are negative. Only in the case of person 32 on the ELTS scale did rater 3 give a much higher band than would have been expected. It will be recalled that the standard deviation of the bands awarded by rater 3 was small, and so it is to be expected that this particular rater is unprepared to use the full width of the band scales, and as a direct result more able students have to perform much better to be awarded a higher band than they would have to be with other raters.

When questioned about this feature of the rating process, rater 3 did demonstrate an overt awareness of the leniency involved in awarding bands, in that students of lower ability could achieve higher bands fairly quickly, while the use of only the middle range of a rating scale resulted in compression of scores and a degree of failure to discriminate (see pages 496 - 498).

However, referring to Tables 8.28, 8.29 and 8.30 in the Appendix to Chapter 8 once more, we may observe that the Mean Square (Mnsq) Outfit statistic for each rater is not above 2 or below - 2. Scores of above 2 are considered to be misfitting across the whole data, while scores of less than - 2 do not contribute unique information. This commonly used Rasch fit statistic is a test of the unidimensionality of the data, in this case rater behaviour without a task variable. Thus, although we have noted inconsistencies in rater behaviour, these
are not serious enough to suggest that the raters are not treating each scale as a unidimensional construct.

4.3 A Partial Credit Model to analyze rating patterns

In this analysis a Partial Credit Model was used, but the task variable was still not introduced. This allows the investigation of the way in which each individual rater interpreted the rating scale thresholds of each individual scale (McNamara and Adams, 1991, 12).

The Item Characteristic Curves for each rater on each scale are reproduced in Figures 8.9, 8.10 and 8.11 on pages 283 - 285. Each curve represents the probability of a student at any given ability level achieving a given band.

In these figures, we can begin to see the variation between raters even given relatively high reliability coefficients. In Figure 8.9 (Fluency) we can see that rater 2 (the most severe) uses bands 2, 3 and 5 to cover the entire ability range. Band 6 is very difficult to achieve, and band 4 cannot be distinguished from band 5. Rater 1 (the second most severe) interpreted band 6 in a similar way, with band 5 being used for most of the more able students. However, this rater was able to distinguish band 4 from other bands. Raters 4 and 5 gave more able students band 6, but at the lower ability levels it was harder for students to gain a band 3. Rater 3 differs from all others in being very lenient with less
Fig. 8.9. Item Characteristic Curves for five raters on the Fluency rating scale.
Fig. 8.10. Item Characteristic Curves for five raters on the Accuracy rating scale
Fig. 8.11. Item Characteristic Curves for five raters on the ELTS global rating scale

Rater 1

Rater 2

Rater 3

Rater 4

Rater 5

285
able students, but also being unwilling to use band 6 for the very able. Band 4 is used for students once they reach an ability of - 1.74 logits.

The picture with regard to the measurement of accuracy in Figure 8.10 is somewhat different. However, it is once again the case that rater 3 awards higher bands to students of lower ability levels with only bands 3 and 4 being used to discriminate between less and more able students. Other raters interpret the scale in a very similar way, although for rater 4 the bands are narrower than for the other raters: it is harder to achieve a band 3 but much easier to achieve a band 5 than with other raters.

Figure 8.11 demonstrates that there is considerable variability in the interpretation of the ELTS global scale. Rater 3 is again more lenient with students of lower ability, but it is harder to achieve higher bands; with rater 2 it is possible to achieve higher bands at lower ability levels, and band 8 is also available at 2.48 logits and band 9 at 4.94 logits.

In the case of each scale, however, there is only likely to be a difference of one band between raters at any ability level (one exception being for students placed at approximately 3 logits on the ELTS scale). This is demonstrated in Tables 8.32, 8.33 and 8.34 in the Appendix to Chapter 8 which present the band which a student could expect to be awarded at a given ability level by rater. For example, in Table 8.32 a student at
around 1 logit would be more likely to be awarded a score of 4 by raters 1, 3, 4 and 5, but a 3 by rater 2. Minor fluctuations such as this when using untrained raters appears to be a strong indication of the reliability of the rating process.

4.4 The effect of task on rating patterns

Once the task is included in the analysis as a test method facet more variability is introduced to the whole picture of reliable oral assessment. The ICCs for each rater on each task (15 diagrams for each scale) are produced in Figures 8.12, 8.13 and 8.14 on pages 288 - 290. By looking down a column one may compare how different raters interpret the rating scale on the same task, whereas by looking at the rows it is possible to see how each individual rater interprets the rating scale on different tasks.

When using the Fluency scale rater 3 is considerably more lenient in the rating of poorer candidates on all tasks, whereas rater 4 is considerably more severe on less able students on Task 3 (Group discussion) only. Task 2 is the hardest on which to score highly, while it was generally easier to score more highly on Task 3.

In the use of the Accuracy scale rater 3 was only particularly lenient to poorer students on Task 2, while rater 4 was only harsh to students of around 0 logit ability on Task 3. Task 2 was again the most difficult on which to gain higher band scores and easiest on Task 3.
Fig. 8.12. Item Characteristic Curves for five raters on the Fluency rating scale taking task variability into account.
Fig. 8.13. Item Characteristic Curves for five raters on the Accuracy rating scale taking task variability into account.
Fig 8.14. Item Characteristic Curves for five raters on the ELTS global rating scale taking task variability into account.
When using the ELTS band scale there was some variability between raters and considerable variability between tasks. Rater 1 was considerably more severe on poorer students on Task 2 than on other tasks, while rater 3 was lenient with poorer students only on Task 3. Rater 4 was somewhat more severe on students of an ability of around - 2 logits only on Task 3. Rater 2 was particularly lenient with better students on Task 3, while rater 5 was particularly severe with more able students on Task 2.

Across raters and tasks the variability which exists is mainly with less able students, who could receive a band of anything between 3 and 6. Rater 1 shows more variability than other raters. This mainly occurs with students with an ability of around - 1 logits, who would have been most likely to receive a band 4 on Task 1, a band 5 on Task 2 and a band 6 on Task 3, and students with an ability of around + 5 logits, who would have been most likely to receive a band 7 on Tasks 1 and 3 and a band 9 on Task 2.

Tables 8.35, 8.36 and 8.37 in the Appendix to Chapter 8 show estimates of judge severity on the Fluency, Accuracy and ELTS rating scales respectively, taking the task variable into account. It can once again be seen that, despite the variability already discussed, no outfit statistics are less than - 2, although we note that in Table 8.35 the outfit statistic for rater 3 is a rather high 2.4 and that in Table 8.37 the outfit
statistic for rater 3 is exactly 2, indicating that once task effect is taken into account the ratings of rater 3 do begin to become much more unreliable.

4.5 Task difficulty

It was hardest to score highly on Task 2, but there was no essential difference between the difficulty levels of the other two tasks.

At this point we may note that task difficulty does appear to be relatively stable. In the difficulty tables (Tables 8.38 - 8.40 in the Appendix to Chapter 8) the higher the logit the more difficult the task, and the lower the logit the easier the task. This information is presented by rating scale. The outfit statistics show that each task was operating unidimensionally, implying that no single task was significantly influencing scores on any of the scales when raters are held constant.

4.6 Partial Credit reliability estimates

Finally, we may turn to Rasch reliability coefficients for the Rasch estimates discussed above. The software used in this analysis provides a reliability statistic for each of the rating scales once the task and rater variables are taken into account. It is the "Rasch equivalent to the KR-20 or Cronbach Alpha" (Linacre and Wright, 1990: 25) and is calculated as true variation divided by observed variation. In other words, it amounts to a test of reliability in terms of unidimensional
structure across the facets isolated (McNamara, 1991),
and these results are presented in Figure 8.15.

Fig. 8.15 Reliability statistics for each rating scale by
raters and tasks

<table>
<thead>
<tr>
<th></th>
<th>Fluency</th>
<th>Accuracy</th>
<th>ELTS global scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters</td>
<td>.94</td>
<td>.65</td>
<td>.91</td>
</tr>
<tr>
<td>Tasks</td>
<td>.91</td>
<td>.83</td>
<td>.59</td>
</tr>
</tbody>
</table>

The Accuracy rating scale is most susceptible to
unreliable variation between raters, whilst the ELTS
rating scale is most susceptible to unreliable variation
between tasks, and it must be accepted that the ELTS
rating scale was not designed to be task independent. The
Fluency scale, however, seems to be used with an adequate
degree of reliability across raters and tasks. That is,
the Fluency scale is not sensitive to task variability
whereas the ELTS scale is. It is argued here that this is
a great strength of the Fluency scale, as reliable data
collection in testing implies that similar results be
obtained irrespective of the method of collection, that
is, the task used. Undue sensitivity to task is the
interference of a test method facet which reduces the
amount of true variance as a proportion of observed
variance, thus reducing reliability.

Section 5: **Intra-rater reliability**

An intra-rater reliability study was conducted
exactly three months after the original rating of all students on all tasks was completed. In this study, all five raters were required to rate 8 of the 47 students for a second time on all scales. In Table 8.41 Pearson Product Moment correlations are reported for all raters across the two separate ratings in the use of each scale, and Table 8.42 reports the Spearman Rank Order correlations. Both tables are contained in the Appendix to Chapter 8.

Finally, Table 8.43 in the Appendix to Chapter 8 presents descriptive statistics for the intra-rater study. In the table, the mean, standard deviation and range are reported for each of the eight students on the first and second rating for comparison.

With the one exception of rater 3, all intra-rater reliability coefficients were satisfactory on the use of all scales. Variations in means and standard deviations are not large enough to cause concern either, except in the case of rater 3. The results indicate that four of the five raters are capable of being reasonably internally consistent, the most important aspect of reliability for Alderson (1991b: 64). The problem of rater 3 will be raised again later.

Section 6: Discussion

6.1 Task 3

(1986) argue strongly that in any oral testing situation it is essential to use a range of tasks, as no single task is capable of eliciting a range of assessable language which would allow the test developer to claim that the procedure was valid. Pollitt and Hutchinson (1987) have also shown the variation which can result from the use of various task types.

With specific regard to the Group Discussion (Task 3) the earliest statement in its favour as opposed to other task types was made by Folland and Robertson (1976). There have been reports of the task type being used successfully in Israel (Reves, 1980; Reves, 1991; Shohamy, 1986) and in Zambia (Hilsdon, 1991) with school aged students, and in Italy with university students (Lombardo, 1984). Berkoff (1985: 95) argues that this task type overcomes the problems of "artificial conversation" between a "distant examiner" and a "nervous examinee". Morrison and Lee (1985) also report successful uses of the Group Discussion with university students in Hong Kong in order to simulate academic tutorials. One study, however, concluded that students could not be fairly assessed in Group discussion tasks, but agreed with other researchers that they require different skills from those needed in oral interviews (Pollitt and Hutchinson, 1987).

If, as is claimed in the literature, the Group Discussion task is so different from other interview task types, the results of this reliability study may not be
surprising. From the G-study it appears that the Group Discussion is significantly different from other tasks. It is important to ask whether this difference is a source of unreliability, or merely reflects the need to include a variety of task types within any oral test as the language elicited and examiners' responses to the language are different.

We may note that inter-rater correlation coefficients were higher for all scores awarded on this task, irrespective of the rating scale being used (pages 276 - 277). Secondly, the task was consistently the easiest on which to gain higher scores. Thirdly, when raters were interviewed regarding task 3, all but one claimed that in their opinion it provided more assessable language and allowed the students to demonstrate with more ease and freedom what they were actually capable of (see page 346). The one rater who did not share this opinion was rater 3. Fourthly, when looking at student questionnaire responses to the various tasks, the Group Discussion was preferred by more students than the other tasks.

It would appear that a variety of tasks are required in any oral test, as the Group Discussion task not only operates in a different way from the other two tasks, it is also perceived in a different way by both raters and students. As it is important that the widest possible sample of assessable language is taken in an oral test, it would seem essential that tasks of this nature are
6.2 Inter- and intra-rater reliability

Most unreliability was caused by rater 3, who was consistently more lenient than other raters and failed to discriminate between less and more able students. Only on Task 3 was this amount of unreliability reduced to some extent. It was discovered that a large degree of agreement did exist across all raters, however, usually within one band on any rating scale.

It was noted, however (page 277) that lower inter-rater correlations were discovered when raters were using the ELTS rating scale. This was more alarming as three trained and experienced ELTS examiners in the study significantly differed in their use of the scale (pages 272 - 273), and there was a larger amount of variability in the use of the scale across tasks (page 293) than in the use of the other scales. It was argued (see pages 86 - 87) that sensitivity to task is not appropriate for a rating scale which is designed to measure underlying ability (communicative or linguistic) which is not task dependent. Task sensitivity to the degree found in the operation of the ELTS scale would only be acceptable in a scale designed to describe performance on a specific type of task, where the criteria for rating would not be sampled in different task types.

It would appear from these results that rater training for ELTS examiners has had little effect upon
rating patterns. This raises the whole issue of examiner training for the oral component of language examinations, dealt with below.

6.3 Implications for rater training

It has become axiomatic in the literature on oral testing that rater training is essential (Wilds, 1975: 34). With some examinations such as the ILR, recertification of examiners is essential every few years (Lowe, 1987) in order to try to ensure that they are marking in line with wider standards. Training and certification is usually achieved through the marking of a sample set of interviews either on video or on audio tape, and a significant degree of agreement between examiners is expected. Adams (1980: 5) makes explicit why most writers on testing are opposed to using untrained raters: naive raters are incapable of seeing or understanding the factors which are important at each level in the scale, and are therefore going to be unreliable in how they interpret the scale thresholds. This approach, however, begs the question of the validity of the scale's representativeness of an underlying ability continuum. That is, by training raters to match examples of performance to numbers (whether they are provided with descriptors or not) one does not know anything about the validity of the scale being used, only something about how reliable the process of matching performance to number can be achieved.
Ingram (1985b: 252) stresses that training is essential, as "The interviewer would seem to be a pivot around which direct testing would ride or fall." Variation in scale interpretation could, according to Ingram, mean that oral assessment would become impossible. Barnwell (1989) demonstrated that 14 untrained raters marking four audio cassettes of learners of Spanish "differed substantially as to how to translate their perceptions into points on the ACTFL scale" (ibid., 158). He concludes that without training low inter-rater reliability is only to be expected. In this particular study it is important to note that Barnwell lays the blame for not achieving inter-rater reliability on the ACTFL scale itself, as the band descriptors are not in themselves self-sufficient or explanatory.

This is an important point. The type of examiner training which is described in Liskin - Gasparro (1984a: 481 - 482) and Duncan (1987), and which is practised by most examination boards, is what Alderson (1991b: 64) refers to as "cloning", as discussed on pages 13, 49 - 50. If untrained raters tend to agree in their interpretation of scale thresholds then it would seem likely that the scale cannot only be used reliably, but that it contains "psychological" validity.

We will return to this in the discussion of validity. However, from the evidence presented here, it would seem more appropriate to develop rater training which is specific to individuals and to tasks (or which
is at least non-directive, as in Knight, 1992). Nevertheless, raters used in this study, with the exception of rater 3, are internally consistent within acceptable and reasonable limits. It would seem to be more appropriate that after rating a number of sample oral tests, Rasch techniques could be used to isolate an individual’s rating pattern across a variety of tasks. Having done this, it may be appropriate to encourage an individual rater to change his or her rating pattern in certain ways which might increase inter-rater reliability. Developments or changes in rating styles or patterns could be monitored through periodic re-rating of a small number of tests and further training provided as and when necessary.

This approach to rater training would allow variation in rating patterns to be taken into account, whilst the use of untrained raters would allow the further investigation of test validity as described above. Further, any rating scale used for any specific task would have to demonstrate that it was robust in use for the developer to claim reliability. This would imply that the rating scale be either (i) specific to a task and used reliably by raters or, (ii) if general (that is, intending to measure trait rather than describe performance) it should be seen to be reliable across tasks and raters. This amounts to nothing more than saying that in the first instance the test method facet is made part of test, and any variance caused by this
method facet is treated as true rather than error variance, whereas in the second instance the test method facet of task is considered to represent error variance.

In this study the Fluency rating scale meets the second of these two criteria. The Accuracy scale is not used reliably in that raters tend to differ in its use to a greater extent than the other scales, while the ELTS scale is not used reliably in that it is susceptible to influence from the test method facet of task. However, bands awarded from the Fluency and Accuracy rating scales have been shown in the G-study to co-vary to a greater degree than those from the ELTS scale. This could tentatively be claimed to be related to the greater degree of detail in the band descriptors and a larger degree of validity which could be attributed to the development process used.

The unreliability associated with the intra- and inter-rater statistics for rater 3 lead to the exclusion of this rater from the validity studies reported in Chapter 9.
Chapter Nine

Validity

Section 1: Introduction

It is accepted that the concept of validity is unitary (Messick, 1989a), even to the extent as suggested in Chapter 8 that reliability studies often impact upon validity. However, for the practical purposes of research it is appropriate to retain the traditional terminology which distinguishes between different types of validity (Messick, 1989b). Terminology such as "concurrent" or "content" validity are taken to be aspects of the superordinate "construct" validity, to which all evidence contributes.

Section 2: Two Concurrent Studies

2.1 Introduction to study one

Of the 47 students who took all three of the tasks developed and were rated on the three scales used in the study, 45 also took the University of Cambridge Local Examinations Syndicate’s First Certificate in English (FCE) Examination within one month of the completion of the study. This provides separate assessments of the students in question in the areas of reading, writing, grammar, listening and speaking. Secondly, the FCE ratings provide separate assessments in the six areas of fluency, accuracy, pronunciation of individual sounds, pronunciation of sentences, interactive communication, and vocabulary (Hamp-Lyons, 1987).
The data collected therefore allow examination of the relationship of speaking measures to those in other skill areas, and the relationship between each of the speaking measures.

2.2 Study One: FCE - Developmental tests concurrent study

2.2.1 The five FCE papers, Fluency, Accuracy and ELTS

In this study it was decided to use the Rasch estimates of student ability on the Fluency, Accuracy and ELTS scales rather than raw scores, as these estimates take into account variation in the test method facets of task and rater as outlined in the reliability study (see page 278). As such, they represent the most accurate estimate of oral proficiency derived from the study. Table 9.1 in the Appendix to Chapter 9 contains correlations between the three rating scales and the five papers of the FCE examination.

Very similar correlations were discovered when the pooled and averaged raw scores were correlated with the five papers in the FCE examination. In each case, the scores from the direct oral interview correlated most highly with the Use of English paper. This is a similar finding to that of Ingram's between the ASLPR and CELT, and the findings of Hinofotis, Bailey and Stern (1981), although in the latter case results are suspect because of the use of stepwise regression techniques. What is perhaps surprising about the correlation matrix is the lower correlation with the listening paper, which one may
have thought would have had a higher correlation with oral measures than reading or writing. It may also be noted that the degree of correlation between both the Fluency and Accuracy scale and the components of the FCE examination is almost identical. This could be an initial indication that the two rating scales developed in this study were not being used independently of each other by the raters.

The correlation matrix in Table 9.1 was factored using iterative Principal Axis Factor Analysis, the results of which are presented in Table 9.2 in the Appendix to Chapter 9.

The results raise the question of whether or not we are faced with a large g-factor accounting for 73% of variance, which some measures tap with greater accuracy than others. However, great caution should be exercised regarding such interpretations (Woods, 1983), as the results of exploratory factor analytical studies of this nature are extremely difficult to interpret. (In this case, the high communalities for the three rating scales may be a result of the fact that they are ratings of the same performance.) In particular, 27% of variance is not accounted for when only factors with an eigenvalue of 1 or more are extracted. The temptation is then to extract two or more factors in order to increase the amount of variance accounted for. The researcher then falls into the trap of fishing for a factor structure which is interpretable in the light of original hypotheses.
(McNemar, 1951). On the other hand, in cases such as this where 27.33% of variance is unaccounted for it is permissible to force the number of factors to be extracted. The problem is when to stop! Farhady (1983: 18-19) for example, argues that:

"...it makes sense to use a cutoff eigenvalue of less than unity because the communalities of interest...are less than unity.... When a judgement must be made concerning a larger or smaller number of initial factors, it seems reasonable to start with the larger number because it is better to examine all the meaningful factors at the start and then eliminate unnecessary ones rather than exclude meaningful factors from the beginning without careful examination. Eliminating any factor without sufficient care may well result in loss of information and distortion of the final outcome."

In this analysis three factors were extracted. Initial eigenvalues were 5.914, 1.174 and 0.394 respectively. The fourth factor had an initial eigenvalue of 0.09 and this was considered too small to justify including a fourth factor. The three factors in the present study were submitted to Varimax Rotation, the results of which are presented in Table 9.3 in the Appendix to Chapter 9.

The listening test alone loads highly on factor 2, whilst all other loadings on factors 1 and 3 appear to be test specific. The FCE speaking test does load more highly on factor 1 than other components of the FCE, and this may be compared with the weaker evidence (an FCE
speaking loading of only .173) upon which Davidson and Bachman (1990: 40) claimed to have isolated an oral trait when comparing the FCE with the SPEAK. Despite this evidence it cannot be entirely discounted that, with the exception of listening, what we observe here is attributable to variance in test method facets rather than traits.

It would seem from using this traditional approach to concurrent validation that there is some external evidence of divergent and convergent validity for the developmental scales of Fluency and Accuracy, but that this evidence is too weak to make great claims.

2.2.2 FCE oral rating components, Fluency, Accuracy and ELTS

A similar procedure was followed in assessing the relationship of the FCE components of the rating scale to the Fluency, Accuracy and ELTS scales to that followed in assessing the relationship of the five FCE papers to the rating scales. (Davidson and Bachman (1990) do not investigate the FCE component rating scales, comparing only the global score to the components of the Speak). In the FCE oral test, examiners are asked to rate students on the six components of Fluency, Grammatical Accuracy, Pronunciation of Sentences, Pronunciation of Individual Sounds, Communicative Interaction and Vocabulary. In Tables 9.4 to 9.9 in the Appendix to Chapter 9 these have been shortened to Fluency, Accuracy, Prosent, Prosound,
Interact and Vocab respectively.

Initially, all the measures were correlated on the assumption that scores on the "Fluency" and "Accuracy" subscales in the two tests may be more highly correlated with each other than with the other subscales, and the results are presented in Table 9.4. The correlation matrix is almost impossible to interpret, but would tend to suggest that separate components of the superordinate "oral ability" were not being validly tapped. That is, effective measurement could be carried out on a global rating scale rather than a componential scale, as there is no evidence for either convergent or divergent validity. It is highly likely that the correlation matrix in Table 9.4 represents an excellent example of contamination across scales in a given administration (Alderson, 1991b). The factor analysis of the matrix in which factors with an eigenvalue of 1 or more are extracted (Table 9.5) bears this out. Once again, the evidence could lead us to the conclusion that a strong g-factor has been discovered.

When two factors were extracted it can be seen that the various scales are operating in significantly different ways, perhaps indicating the influence of a task - or more likely an "examination" - effect. Initial eigenvalues for the two factors extracted were 6.91 and 0.819 respectively. The third, at 0.07 was considered too small to include. The results of the two factor solution are presented in Table 9.6 in the Appendix to Chapter 9.
This may be a sign that although the measurement of a separate "oral trait" is valid, it is not valid, given the current state of knowledge about what constitutes that oral trait, to break it down into component parts. However, it is unwise to draw such conclusions from a concurrent study, and other methods of analysis must be called upon to further investigate the nature of the component scales.

2.2.3 Discussion

Limited evidence has been found to suggest that an oral trait can be identified from the concurrent study, but no evidence to suggest that this oral trait can be broken down into valid components. However, this is not to be considered surprising in a concurrent study.

Firstly, no data is available on the reliability of the FCE papers. Essentially, this study has had to assume that reliability of the FCE is similar to that for each of the scales with which they were correlated. If any of the criterion measures are unreliable this in itself would affect the results of the concurrent study. It is noted that Davidson and Bachman (1990: 35) found that reliability estimates for the FCE reading paper of .79 and listening paper of .62 were "below normally accepted test development standards." Reliabilities were not available for writing or speaking, while the Use of English Paper had a reliability coefficient of .85. If this were true of the FCE papers taken by the 45 students
used in this study, this in itself could account for the correlations in Table 9.1 in the Appendix to Chapter 9, and the apparent link between the three oral scales and the Use of English paper.

Secondly, when considering the relationship between the Fluency and Accuracy rating components of the FCE oral test with the developmental Fluency and Accuracy scales it has to be assumed that, for example, the FCE Fluency scale and the developmental Fluency scale do actually tap the same trait. This refers to the requirement that the criterion or the predictor must be construct valid in order to draw conclusions. Even a cursory glance at the descriptors in the two scales (see pages 123 and 160 - 163) would tend to indicate that the construct of "Fluency" is interpreted in different ways. If this is the case, the finding that the two are unrelated is not terribly surprising.

Thirdly, it has been discovered that there is a task variable and a rater variable which affect scores and are a significant source of test variance (pages 268 - 273). The tasks taken by the students in the FCE oral were varied, and each student took only one task with one rater. This means that neither the task variable nor the rater variable can be controlled for in the criterion. What this amounts to is that we know nothing of the effects of test method facets in the criterion.

This raises once again the major problem in conducting criterion validation studies with oral tests:
selecting an appropriate criterion. It would seem that
the only way to do this would be to develop a number of
fluency scales independently of each other but from the
same specification of the construct, in a similar way to
which item writers produce individual test items from
specifications. The separate fluency scales could then
reasonably be analyzed to discover if they were "parallel
operationalizations" of the same construct. This
criticism applies not only to the study which has been
carried out here, but to all the research which has
recently been conducted between OPIs and SOPIs.

It would therefore seem appropriate to conclude that
concurrent studies to compare component rating scales of
oral ability are rather premature, as the issues are so
complex and the factors involved which may affect scores
so little understood that inconclusive evidence should
not be surprising in a concurrent study.

2.3 Study Two: The relationship between teacher
assessments of student oral abilities and the
Developmental scales

2.3.1 Introduction

The literature on the relationship between test
scores and judgements of ability is concerned either with
predictive validity, that is, whether test scores can
predict the assessment of student performance in some
future non-test situation (Messick, 1989a: 71 - 72), or
with concurrent investigation into the relationship
between tests taken within a limited period of time. In
the case of predictive validity the person who assesses the student in the non-test situation is usually not a language teacher or tester, but someone from another discipline in which the student is working.

One such predictive study (Criper and Davies, 1988: 54 - 55) investigated the relationship between ELTS scores and academic supervisors' judgements of student language ability at the beginning and end of their academic courses. At the beginning of the course the correlation between the ELTS speaking scores and the supervisors' assessments of student speaking ability was .43, and at the end of the course .42. The concurrent study between the ELTS oral score and language tutors' assessments of students on a professional/academic course also produced a correlation of .42 (ibid., 56). Even if there were problems with the way in which the data for the Criper and Davies studies were collected (Rea and Wall, 1988: 56 - 57) it is very hard to believe that these correlations would have been much higher.

A concurrent study was carried out by Levine, Haus and Cort (1987) to investigate the assessment of students by teachers of French and Spanish on the ACTFL scale in comparison to the students' actual ACTFL results. Using t-tests rather than correlations they discovered that teacher assessments were significantly different from ACTFL ratings at p < .01 (ibid., 47). Levine et. al. came to the conclusion that teachers were in need of training on the use of the ACTFL scale in the classroom in order
to make them more accurate assessors of their students' ability levels - the return to the problem of "cloning" (Alderson, 1991b).

Here, we are not concerned with issues of predictive validity, but with the extent to which oral test scores are associated with concurrent assessment of oral ability made by teachers. Although Levine et. al. assume that concurrent teacher assessment and oral test scores should coincide, their own data show that this is not the case, as does the data of Fulcher (1991). The difference between teacher assessment and oral test scores should not be seen as surprising, as a classroom teacher working with a student over a lengthy period of time has the opportunity to form an opinion regarding the student's oral ability in a wide variety of activities and on a large number of tasks. On the other hand, assessing oral ability unsystematically in the classroom also provides the teachers with opportunities to make errors which may not be present in more controlled testing situations. In the oral test a much smaller number of tasks are also used in a very short period of time. A completely different set of facets are involved in the two situations, and have not been sufficiently investigated in the classroom situation. The teacher hopes that her assessment of a student is reasonably accurate; the oral test developer hopes that the evidence collected in the test will allow the rater to assign a reliable and valid score. But there is no a-priori reason why the
differences between the two assessment situations should not produce different results.

It may very well be that as much more research is done into the way in which teachers assess their own students in the classroom setting, it would become possible to combine teacher assessments with the results of oral tests to provide information which is more valid.

Once again, problems with criterion studies are raised. When Criper and Davies (1988: 55) state that "...discrepancies between supervisors' judgements and test results become themselves an interesting area for investigation..." they are perfectly correct. Transferring this to language teachers' concurrent assessments we have to say that very little is known about the way in which teachers make assessments and how they collect and synthesise their evidence. As a result it is impossible to control for classroom assessment facets in relation to test method facets.

2.3.2 Available data from teacher assessments

In the introduction it has been argued that teacher assessments cannot presently be used in test validation studies. However, as this type of concurrent study is considered to be one aspect of validity which should be investigated, appropriate data were collected in order to see whether or not the correlations would be in the order of .5 to .7 which appears to be usual for teacher assessment and external measures (Fulcher, 1991).
Within one month of the students taking the oral tests developed for this study, their teachers were asked to provide assessments in the areas of speaking, listening, reading, writing, grammatical accuracy, spelling, and the quantity and quality of work produced outside class during the year. This information was collected on a standard form which uses a Likert-type scale for each category, on which the teacher is asked to tick one of five boxes headed "very good", "good", "adequate", "poor" and "very poor" for every student. The assessments were analyzed using the same procedures as in the concurrent study with the FCE examination results, and the results are presented in Tables 9.7 to 9.10 in the Appendix to Chapter 9.

2.3.3 Analysis of concurrent teacher assessments

It can be seen in Table 9.9 in the Appendix to Chapter 9 that a one factor solution can account for the data collected on teacher assessment and the rating scales developed for the study. However, a four factor solution was also forced, in which the initial eigenvalues for the first four factors were 7.66, 0.69, 0.41 and 0.35 respectively. The fifth factor was not included as the eigenvalue was estimated at 0.06. Table 9.10 in the Appendix to Chapter 9 shows that ratings of one performance will tend to load highly on the same factor. Although all the oral scales load on factor 2, the assessment of speaking by teachers loads on factor 4.
Section 3: Validity by Group Differences

A fairly weak form of validity (in the case of second/foreign language testing), is that of group differences (Messick, 1989a: 55). The principle behind the analysis of group differences is that different sections of the population taking a test may be identified in advance of taking the test. The test results may then be described as valid if they do in fact discriminate between the various groups identified in advance.

In the case of the students used in the current research, they were identified as coming from three distinct sub-groups of the population. Group 1 were identified as "good language learners," Group 2 were "average language learners" and Group 3 "poor language learners" (see pages 242 - 243). This was done on the basis of their language learning history and reports from language teachers.

For the purpose of this kind of validity, it would be expected that the use of a valid rating scale would separate these students out into the appropriate group, even though the raters could not have known into which group they had been previously classified.

For the purposes of this analysis the Rasch scores on Fluency, Accuracy and ELTS were used. In order to test whether or not ratings on each of the scales did discriminate between students One Way Analysis of Variance was used, together with a post-hoc Tukey HSD
In the Appendix to Chapter 9, Table 9.11 contains the results for the Fluency scale, Table 9.12 for the Accuracy scale, and Table 9.13 for the ELTS scale. It will be seen that in each case the between groups variance is larger than within groups variance, and that the differences between groups 1 and 2, 2 and 3 (and hence 1 and 3) are indeed all significant at $p = .01$ or less. It should be noted that in this analysis only 45 students were included in the results, as one student was originally removed from the sample due to the poor quality of the video recording (see page 245) and one student was taken ill immediately after the experiment had been concluded and therefore did not complete the post-experiment questionnaire. Summary statistics for the groups are given in Table 9.14 in the Appendix to Chapter 9.

In the rather gross terms of validity by Group Differences, it would appear that each of the scales possess group discrimination validity. However, this form of validity is weak, and cannot be used to claim that the rating scales themselves, or the band descriptors, contain anything which relates to psychological reality. For this, much more precise tools are required.
Section 4: **Construct Validity**

4.1 **A Multitrait - Multimethod study**

The validity of the scales developed was further analyzed using three separate methods. Initially, a Multitrait-Multimethod (MTMM) study was conducted. In the MTMM study average correlation coefficients from four raters were used. The three test method facets are the three oral tasks, and the three traits were those of Fluency, Accuracy and ELTS global ratings. Table 9.15 in the Appendix to Chapter 9 presents the MTMM matrix for the uncorrected correlation coefficients. From an initial inspection of Table 9.15 we may immediately note that all the correlation coefficients are high, indicating that it may not be realistic to expect evidence of divergent validity between the ratings of Fluency and Accuracy, and making the investigation of the matrix corrected for attenuation a pointless exercise. It must also be borne in mind that the high correlations suggest that the MTMM approach to investigating divergent validity may not be entirely appropriate for this data set, even if some light is thrown upon its structure.

The requirements for MTMM studies (Campbell and Fiske, 1959) are very hard to meet. The method requires that the correlation matrix demonstrates convergent validity between the same trait across methods, and divergent validity between distinct traits irrespective of the method used to measure those traits.

The first requirement is that "the entries in the
validity diagonal should be significantly different from zero and sufficiently large to encourage further examination of validity." (ibid., 82). In Table 9.15 this requirement is met, providing some evidence for convergent validity. The second requirement is that "a validity diagonal value should be higher than the values lying in its column and row in the heterotrait-heteromethod triangles" (ibid., 82), which would indicate that the variable shows divergent validity from variables which have neither method nor trait in common.

The validity diagonals singularly fail in this respect, suggesting serious invalidity due to the influence of factors other than trait in the scores. However, it is necessary to add that the correlations are so high in both cases as to make it nearly impossible to meet this requirement.

The third requirement is that "a variable correlate higher with an independent effort to measure the same trait than with measures designed to get at different traits which happen to employ the same method" (ibid., 83), which would demonstrate that a method factor was not significantly influencing results. The correlation coefficients in the heterotrait-monomethod triangles are consistently higher than all entries in the validity diagonal in Table 9.15, indicating a strong method effect, which would suggest that the measures of Fluency and Accuracy are not valid as measures which can be used without consideration of the method facet of task.
The final requirement is that "the same pattern of trait interrelationship be shown in all of the heterotrait triangles of both the monomethod and heteromethod blocks." (ibid., 83) Although this is certainly the case with the matrix in Table 9.15, this cannot be taken as evidence of validity as all correlation coefficients are so high as to be uninterpretable. It would appear from the first analysis of this data that although "oral ability" may be being measured, it is not possible to break this down into valid components. There are two possible explanations of these findings. Either the results confirm those of the criterion study, namely that what we have is a "halo effect" from the methods, or alternatively all the methods employed tap the same unitary trait.

The MTMM approach to construct validation requires that any hypothesized trait be distinct from methods. What this means is that the trait can be measured under a variety of experimental conditions for it to be considered conceptually sound. Campbell and Fiske (ibid., 100) put the issue very clearly:

"...any conceptual formulation of trait will usually include implicitly the proposition that this trait is a response tendency which can be observed under more than one experimental condition and that this trait can be meaningfully differentiated from other traits."

What this amounts to is that any hypothesized trait which is essentially method dependent is not actually a
trait. The initial evidence from the MTMM study of the rating scales of Fluency and Accuracy which have been designed as reflecting two separate traits, suggests that they do not tap the hypothesized traits, and therefore lack construct validity as originally defined and developed. However, the inadequacy of the MTMM approach to validation given such high correlations requires that other methods are used to analyze the data.

4.2 Maximum Likelihood Confirmatory Factor Analysis

As in Bachman and Palmer (1981, 1982, 1983) the MTMM study was followed up with a Maximum Likelihood Factor Analysis (ML) (Joreskog, 1969) using LISREL 7.16 (Joreskog and Sorbom, 1989). In this study two models were analyzed: a general factor and two partially correlated specific factors, and three correlated method facets. It had been hypothesized that, if the two rating scales of Fluency and Accuracy actually tapped those traits, then the general factor and two partially correlated specific factors would have been the best fitting model.

LISREL 7.16 will not proceed with an analysis if the initial estimates of the fit of the model to the data suggest that the model is grossly inadequate. This occurred with the general factor and two partially correlated specific factors of Fluency and Accuracy. However, the fit statistics produced by the three intercorrelated test method facets were good. Figure 9.1

320
shows the path diagram for this solution.

Fig 9.1. Path Diagram for an ML solution of three intercorrelated test method facets

Squared multiple correlations between each of the X variables and the test method facets (ξ) are used as a measure of their individual relationships, and a total coefficient of determination is a measure of the joint relationship of the X variables to their ξ variables. These are presented in Table 9.16 in the Appendix to Chapter 9.

As all the correlations are high, all the X variables are good measures of the ξ variables. Of course, in this case, this result means that there is a strong test method effect in the scores, confirming the invalidity of the traits suggested in the MTMM study. The goodness of fit of this model to the data is presented in Table 9.17 in the Appendix to Chapter 9.

The $X^2$ statistic is not significant for the correlation matrix, indicating that the model does fit
the data. The Root Mean Square Residual is very low, indicating that the model accounts for a large proportion of the variance in the data. We must therefore conclude that test method facets do account for the small differences which exist between the scores on the three rating scales under investigation. It should be stressed, however, that these differences are extremely small.

The model which was fitted to the data, as made clear in the path diagram, allowed for intercorrelation of the three test method facets. This intercorrelation can be assessed by analyzing the $\Phi$ matrix, which is presented in Table 9.18 in the Appendix to Chapter 9.

The high values in the $\Phi$ matrix suggest that although test method facets do have an effect on scores, these might not be as large as the construct validation results tend to suggest. If the main conclusion to be drawn from the results is that there is cross-contamination on the scales as suggested in the MTMM study, then the ML analysis would only isolate small differences among the test method facets as the main source of variability in scores. Indeed, this possibility is suggested by the factor loadings of the $X$ variables on the $\xi$ variables, where the loadings are generally fairly low. These are presented in Table 9.19 in the Appendix to Chapter 9.

The final assessment of this ML solution is to look at modification indices for $\Lambda X$. The modification indices for the parameters of the model specified in the path
diagram (Figure 9.1.) provide information regarding the potential change in the $X^2$ fit statistic if a previously fixed parameter is freed. The modification index is provided in Table 9.20 in the Appendix to Chapter 9.

The maximum modification index for the matrix is for $\lambda(8,1)$ at 6.06 which is the relationship between the ELTS score on task 2 and task 1. Although freeing this parameter would increase the value of the model fit statistics, there is no theoretical justification for doing this at all. Further, a reduction in $X^2$ by such a small amount would not be particularly useful in the light of the loss of explanatory power that the model currently has.

It can only be concluded from the ML study that it is not possible to provide evidence of construct validity for the rating scales of Fluency and Accuracy. Evidence suggests that the only latent variables at work are those of test method facets, although these do not appear to be exceptionally strong. A second possibility is that each of the scales equally tap a single trait, but that it is not possible to isolate this trait by convergent and divergent criteria. A third possibility is that cross-contamination in the rating process has masked any real differences which might exist between traits. In order to further investigate this, we will turn to Rasch Partial Credit analysis.
4.3 Rasch Partial Credit analysis

McNamara and Adams (1991, 12) conducted one study using Rasch Partial Credit in which the Rating Scale Model was used for judges and the full Partial Credit model used for rating categories. By introducing a full Partial Credit model for tasks (test method facets) as well as rating categories, Rasch analysis can be used not only to investigate reliability but also validity. The Item Characteristic Curves (ICCs) produced by such an analysis on this data produces a 3 x 3 figure in which we may compare the probability of any student with any rater being awarded a particular score on each specific scale when attempting each specific task.

The advantages of this method when dealing with oral testing in particular are great. Whereas the MTMM and ML methods of estimating construct validity only use correlation coefficients as data, the Rasch Partial Credit method uses all the data available from all facets of the testing procedure.

The analysis of construct validity would involve comparing the ICCs across rows and columns. In Figure 9.2 on page 325 the tasks are presented across rows and the three rating scales in the columns. Evidence for construct validity would be available if and only if the ICCs in the columns (rating scales) are very similar, indicating that task does not have a very large influence upon the interpretation of the scale or its individual band descriptors.
Fig. 9.2 Item Characteristic Curves for the study of rating scale validity

Task 1

Task 2

Task 3

ELTS

Accuracy

Fluency
The main problem in interpreting Figure 9.2 is that it is not always very clear where the differences lie between the interpretation of band boundaries across tasks. However, at first glance it appears that there is a large degree of similarity in columns. In order to overcome this problem, numerical data will be analyzed separately, and an attempt will be made to devise a new validity coefficient for use with Rasch Partial Credit models, which can be assigned to each band in each scale on each task.

In the Partial Credit model the probability \( p \) of being awarded a particular band on a rating scale is defined as:

\[
p = \frac{\exp(\beta_n - \delta_{ik})}{1 + \exp(\beta_n - \delta_{ik})}
\]

where \( \beta_n \) is the ability of person \( n \), and \( \delta_{ik} \) is the difficulty of scoring band \( k \) on scale \( i \) (Wright and Masters, 1982: 42).

The statistic \( \delta_{ik} \) can therefore be set out in tabular format for comparison. Again, the hypothesis would be that there would be greater similarity between the \( \delta_{ik} \) statistic for rating scales than for tasks. The statistics are presented in Table 9.21 in the Appendix to Chapter 9.

Even a cursory glance at Table 9.21 reveals that there is more consistency in the columns of the Fluency
scale than the Accuracy scale, and the Accuracy scale
than the ELTS scale. However, from the δik statistic, it
is possible to produce a new validity statistic which
takes all available data into account. The first step in
this process is to produce a new table similar to that of
Table 9.21, in which the entries are deviations from the
mean δik. These are presented in Table 9.22 in the
Appendix to Chapter 9.

The assumption in producing Table 9.22 is that the
mean of δik across tasks within each scale (or δ.κ -
delta dot kappa) would be the best estimate of the true
δik for any given band on the rating scale once the test
method facet of task has been removed. This is not an
unreasonable assumption, as the mean of multiple
observations of the same phenomenon is always assumed to
be statistically more accurate than any single
observation. If, as has been demonstrated in this case,
the measurement is accurate, deviations from the mean can
be attributed to test method facets.

It should be noticed from Table 9.22 that all
columns add to a total of zero, give or take .01 for
errors of rounding to two decimal places. Before
providing a more complete justification for the following
statistic, it is now possible to define a new validity
statistic for Partial Credit models as:

\[
\text{Validity} = 1 - \frac{(\delta.\kappa)^2}{1 + (\delta.\kappa)^2}
\]
This validity statistic uses the deviation scores and plots them onto the validity curve shown in Figure 9.3 on page 330. For example, the lower threshold for achieving a band 4 on the Fluency rating scale for tasks 1, 2 and 3 respectively is 0.24, 0.11 and 0.07 logits. The mean of these three figures is 0.14 logits. δK is the deviation of the lower threshold for the band on each scale from this mean, providing the figures .10, -.03 and -.07 logits. δK is then used in the formula on page 327 to provide a misfit statistic for each instance of a lower threshold as explained above.

This definition will allow the researcher to assign a validity statistic not to a rating scale as a whole, but to each individual band within a rating scale on each individual task, thus taking the use of the Rasch Partial Credit model well beyond uses currently envisaged for the validation of oral rating scales. The validity coefficients derived from this method are always on a scale of 0 (less validity) to 1 (maximum validity), and as such the two extreme points of the scale are absolute. However, between 0 and 1 the validity coefficients which are produced are dependent on the method of calculation, and hence relative. It would, for example, be possible to use a formula which produced a straight line or indeed a curve of a different shape from that in Figure 9.3. The justification for using this particular curve is that all reliability and validity coefficients to date use a
similar curve, although no justification for what happens between 0 and 1 is presented in the literature. Using Rasch Partial Credit models it would be possible to produce an absolute, rather than a relative, interpretation of the validity curve in Figure 9.3, and this will be discussed in Chapter 11.

With this warning in mind, when we look at the validity curve, what we are essentially doing is assessing the degree to which each use of a band on a rating scale on each task is likely to deviate from the best available assessment of the true δik. That is, the deviation from the true point at which a student of the same ability level (θn = δik) would be as likely to score the band below as the band above (p = .5). Thus, given the curve which has been used in Figure 9.3 a high validity coefficient for a the lower threshold of a given band on a given task is an indication of the validity of the band descriptor for the task in question, but not for validity across tasks and hence for generalization of results. High, consistent validity coefficients must be entered in columns for the researcher to make claims of validity for the bands on a particular scale, as this would show that raters are consistently using the band descriptors to classify the same students into the same bands on the rating scale, irrespective of test method facets. Using the above formula, Rasch Partial Credit Validity statistics were generated for each scale by band.
Fig. 9.3 A Rasch Partial Credit validity curve for use with bands in rating scales
and by task, and these are presented in Table 9.23 in the Appendix to Chapter 9.

The validity coefficients for the Fluency scale may be used to illustrate this. The validity coefficients for band 2 on tasks 1, 2 and 3 respectively, are .98, .95 and .94. These coefficients are high and similar across tasks. This may be compared with the same figures for band 2 on the Accuracy scale, where the coefficients are .46, .52 and .91 respectively. These figures for the Accuracy scale indicate that although the band descriptor for the scale appears to be valid for lower ability students on task 3, it is probably not valid for use on tasks 1 and 2. Band 2 cannot therefore be said to be valid, as it does not meet the basic requirement for a measurement instrument that similar results are obtained irrespective of the method of data collection.

The validity coefficients are generally acceptable for the Fluency scale, and highest when raters are assessing students who will get bands 2 - 4. As student ability improves, the scale becomes less valid in assessing them. Although there is variation across tasks, this is only large at bands 5 and 6 (and band 6 was the extreme on the rating scale which was not generated from the data analysis, but added on with band 0 to ensure the use of bands 1 - 5 by raters). The validity coefficients for the Accuracy scale and the ELTS scale fluctuate wildly by both band and task, with the ELTS scale validity coefficients falling to unacceptably low levels.
in many cases.

It is believed that the development of this new validity statistic which allows the analysis of bands within rating scales rather than entire rating scales, is a significant step forward in the process of evaluating oral testing instruments.

4.4 Further Validity Considerations

Having attempted to define and measure one aspect of validity which concerns accurate measurement across test method facets when dealing with individual bands, it is now possible to turn to the issue of the adequacy of the continuum of the scale. This aspect of validity is concerned with the extent to which sequential band descriptors can be said to represent an underlying proficiency continuum from less to more proficient.

The Fluency scale, for example, has seven band descriptors. The assumption is that $0 < 1 < 2 < 3 < 4 < 5 < 6$. Masters (1990, 58) draws attention to this aspect of Fluency as an integral component of original definitions of criterion referencing (Glaser, 1963) which, he claims, has often been overlooked in much modern work on criterion referencing. Masters recommends using ICCs to check whether or not a given rating scale is a continuum. Using ICCs as Masters recommends in conjunction with the rows of the validity coefficients devised above, would indeed provide evidence for or against continuum validity of a rating scale. For example, band 4 of the Accuracy
scale on task 1 is almost subsumed within band 3 (see Figure 9.2 on page 325). This would indicate that the band has no divergent validity from band 3, and does not stand on a continuum. On tasks 2 and 3 we observe that band 3 can be clearly distinguished, but it should also be noticed that the validity coefficient for band 4 on task 2 is only .61, and on task three, the coefficient for band 3 is only .65 (see Table 9.23 in the Appendix to Chapter 9).

The argument here is that the item characteristic curves for successive bands within a scale can be taken to represent the actual performance of a scale only when successive horizontal validity coefficients are high (that is, when read across rows of the table). When this is the case, the item characteristic curves provide evidence of validity when bands are clearly distinguished from the band below and the band above, and preferably when the band widths are reasonably equally spaced across the ability range. Exceptionally wide or narrow bands may indicate a lack of definitional adequacy within the band descriptors. In the former case the band descriptors may not be capable of distinguishing between a range of students at different ability levels (the band descriptors may be too general), whilst in the latter case the band may demonstrate a tendency to be subsumed by other bands (the band descriptors may be too specific).

Returning to the Fluency scale, we may notice that
on tasks 2 and 3, band 4 is narrower than the other bands. This is probably an accurate picture, as on Task 1 the validity coefficient is a slightly lower .90, meaning that we can be less certain about its true positioning on the graph. Thus, although band 4 does appear to be valid, this could be a candidate for careful revision of the descriptor in order to make it slightly more general, thus widening its use. Similarly, band 5 appears to be far too wide on Tasks 1 and 3 which have reasonable validity coefficients. This would be an indication that the descriptors for band 5 need to be more specific in order to make the band applicable to a smaller number of students.

It is claimed here that this approach to validity not only allows the researcher to investigate individual bands within a scale, but the nature of the continuum underlying the scale. Further, the validity coefficients in conjunction with the ICCs provide a method of deciding which band descriptors are valid, and how they may be revised to improve validity.

4.5 Discussion and Conclusions

This approach to construct validity provides reasonable evidence for the validity of the Fluency scale, but does not provide evidence for the construct validity of the Accuracy scale or the ELTS scale. That is, the validity coefficients on the Fluency scale do not vary by column to any great extent, and although there is
some variation by row, an indication of how band descriptors may be changed to improve validity has been discovered and expounded. With the other two scales there is a great deal of random fluctuation by both row and column, showing that more serious revision - probably even complete redrafting - may be necessary.

Although the MTMM and the ML study showed that construct validity was absent from the rating scales, the Rasch Partial Credit model analysis does suggest that the Fluency scale does have a degree of construct validity, especially for students at certain ability levels. This difference in results between the methodologies may not be as surprising as it appears at first sight. Groot (1990, 21) points out quite clearly that MTMM type studies only provide "evidence for insufficient validity of at least one of each pair of tests." If only one rating scale from three used in this study possesses a degree of validity, then this cannot be isolated by an MTMM or ML methods. However, the Rasch approach is capable of testing where the lack of validity pointed up by the MTMM and ML studies originates.

Although the claim for the construct validity of the Fluency scale is based upon the evidence from the Rasch Partial Credit model analysis only, Groot's observation may give more confidence in the results obtained, and to the conclusion that the Fluency scale is not as susceptible to changes in test method facets and more sensitive to changes in student ability level, and that
sequential band descriptors do appear to represent a proficiency continuum.

With this observation, we return finally to the issue of the concept of construct validity. Any hypothesized construct, or trait, depends essentially for its identification and validation on an operationalization which can be shown to be independent of the conditions under which the observations were made and represent an underlying continuum of ability. These are basic criteria of establishing a scientific approach to the measurement of proficiency on rating scales. They are also the difference between measuring student competence and assessing student performance. In the former case there is evidence which will allow generalization to other testing conditions (and, possibly, to situations external to the testing conditions) and comparison with students who are "better" and "poorer", in the latter one may only describe performance under the particular conditions under which the test was held. This relates to the fundamental claim by Bachman and Savignon (1986) that all language tests are indirect in that they attempt to measure a trait or underlying ability which is observed through performance. Rating scales which include references to the testing situation, such as the ACTFL, refer directly to performance factors which inevitably confounds test method facets and traits to such an extent that validation becomes impossible.
This is closely related to Alderson's distinction between the "reporting function" and the "rating function" of a scale (Alderson, 1991a). The reporting function of a scale must contain descriptors which adequately reflect the degree to which test results may be generalized to situations beyond those of the test, and may include references to what would otherwise be classed as test method facets. However, the scale which is used for the function of rating must not contain such references, for their inclusion precludes the possibility of investigating construct validity, and it is only through establishing a degree of construct validity, that a case can be made for the claim that the ability of the student observed under test conditions could be transferred to other situations and other tasks. It may, therefore, be possible to rate on one scale which is linguistic in nature, and report on another which predicts the variety of situations to which one would expect the competence measured to be transferable. The latter scale could be phrased in terminology which would be understandable and acceptable to test result users.

The use of the linguistically valid scale for rating and a separate scale for reporting would seem to be appropriate practice. Language testing in general and oral language testing in particular, does appear to have suffered from the problem of the lack of ability to generalize. It is only through the investigation of construct validity that generalization may take place,
and that a test developer or researcher may claim to have achieved a reasonable degree of validity in any results obtained. For the present, it is with this aspect of test development that advances can be made. The alternative is to admit to relativity and an inability to do other than comment on specific events as they happen in isolation.

In conclusion, it is being argued that sufficient evidence for the validity of bands 2, 3 and 4 of the Fluency rating scale has been discovered to warrant further study of those bands. Band 5 does not appear to possess the qualities needed for validity as described in this section, whilst bands 1 and 6 did not have descriptors but were merely markers of the beginning and end of the scale.

It cannot therefore be said that the use of databased or "data-driven" construction techniques in producing the two developmental scales of Fluency and Accuracy has been an unqualified success. However, more success was clearly enjoyed in the development of the Fluency scale, and this would seem to be in line with the original predictions made from the discriminant analysis results (page 152) during its development. The implications of these findings will be discussed in the conclusions in Chapter 11.
Chapter Ten

Retrospective Report Study

Section 1: Introduction

A retrospective study was conducted in order to gain additional information on the perceptions of raters and more reflective students of the oral interview procedures used in this study. The interview method used was chosen in order to allow the raters and a sample of students to express their views freely on the process, and to allow a reasonable degree of probing of their perceptions when appropriate.

The data presented in this chapter are not "introspective," but what Grotjahn (1986, 167) refers to as a "retrospective verbal report." In the first case, respondents provide spoken data on the way in which they complete a given task whilst they are doing it. In the second case, after the task is completed they are asked to comment on the way in which they completed it. This retrospective debriefing is considered inferior, as it does not provide on-line access to immediate responses which are taken to directly represent mental processes, but allows time for hypothesizing and theory building on the part of the respondent. However, in this research it did not seem possible to get students to comment verbally on an oral test while they were actually doing the test. The five raters were asked to rate a total of 141 video recorded oral tasks totalling some 40 hours. It was clearly not possible to sit with them while they were
doing this, and so they were asked to keep notes on (a) their reactions to the tasks, (b) their reactions to the scales, and (c) any problems they ran into while rating any of the students. They were informed that there would be a thirty minute debriefing session after all the subjects had been rated. Although questions for raters dealt with tasks and general impressions of the testing process, in what follows we will focus upon only those comments to questions which relate specifically to rating scales.

In work on questioning/debriefing techniques outside the field of Applied Linguistics and Testing it is generally accepted that the data collected from an interview in which one is primarily interested in the interviewees' views, are subject to error related to the interviewer and the elicitation procedure (Fowler and Mangione, 1990). Work in "standardized surveys have wide applicability to other types of interviewing" (ibid., 18). In order to ensure that responses were as comparable as possible, a number of rules were followed by the interviewer.

Firstly, one of the major threats to standardized interviewing is not reading the question exactly as worded. Bradburn and Sudman (1979) discovered that more experienced interviewers tend to adapt questions to the context of the interview, but that this process introduced real bias into responses. In this study the questions as worded prior to the interview were asked in
each case in exactly the same way.

Secondly, when responses to questions were not adequate, it was difficult to know when to probe the response in the hope that one would get more interesting data, and when to continue with the next question. If probing was attempted, then an attempt was made to make this non-directive. Fowler and Mangione (op. cit., 42) suggest one of three ways of probing which will tend not to affect the responses gained: "How do you mean?", "Tell me more," and "Anything else?" Unfortunately, in an attempt to be more personable, interviewers do tend to influence the response patterns of certain interviewees, which constitutes interviewer related error. Some probing of answers was attempted in this study, but it was avoided as much as possible especially with the students where (a) they showed that they were not willing or able to provide further useful information, and (b) where time available prevented further probing.

Section 2: Interviews with Raters

Although one or more questions were omitted for some raters who had already dealt with the question under another section of the interview, the questions that follow were read as worded, and any probes used were as non-directive as possible. The responses to these questions may be found in Appendix 10.1 for all five raters. All interviews with raters took place within three weeks of their completing the ratings.
The questions were placed into three blocks: Section A on the tasks, Section B on the rating scales and Section C general questions.

Section A: Tasks

Q1: What kind of language do you think each of the three tasks would have elicited?

Q2: Which of the three tasks did you think was the most difficult/easiest for the students?

Q3: Which task produced the most assessable language in terms of (i) quantity and (ii) quality?

Q4: Did you think any of the topics were unfair?

Q5: If you could revise any of the three tasks, what would you change to make them better?

Q6: Do you have any other comments on the tasks?

Section B: Rating Scales

Q7: Did you find the Fluency scale easy to use?

Q8: Did the Fluency scale contain any descriptors/parts of descriptors which you think are not a part of the concept of fluency?

Q9: Did the Fluency scale not contain anything which is, in your opinion, part of the concept of fluency?

Q10: When rating, did you rely on any criteria other than those in the rating scale?

Q11: Did you find that you used the entire range of possible bands?

Q12: Did you think that the amount of detail in the Fluency scale was appropriate?

Q13: Do you have any other comments on the use of the Fluency scale or descriptors?

Q14: Did you find the Accuracy scale easy to use?
Q15: Did the Accuracy scale contain any descriptors/parts of descriptors which you think are not a part of the concept of Accuracy?

Q16: Did the Accuracy scale not contain anything which is, in your opinion, part of the concept of accuracy?

Q17: When rating, did you rely on any criteria other than those in the rating scale?

Q18: Did you find that you used the entire range of possible bands?

Q19: Did you think that the amount of detail in the Accuracy scale was appropriate?

Q20: Do you have any other comments on the use of the Accuracy scale or descriptors?

Section 3: General Questions

Q21: Were there any particular students who you had difficulty rating either in terms of fluency or accuracy? If so, please explain why?

Q22: Do you think you are a particularly lenient or harsh rater?

Q23: Do you have any further general comments regarding the process of assessing oral proficiency?

Section 3: Interviews with students

Eight students were selected at random from the higher ability group. They were asked to watch their performance on the Picture description task four days after they had completed all of the interviews. As this took approximately 15 minutes per student, the amount of time available to conduct a debriefing interview was limited to only ten to fifteen minutes.

The interviews were recorded and transcribed. The responses for each of the eight students may be found in 343
Appendix 10.2. Only three questions were asked, although students were given more freedom in the amount of detail they went into in response to each question. In some cases, there were preliminary comments from the students immediately after seeing the video of themselves taking the task. As the students could not comment on the rating scales, questions focused on affective responses to the test. The responses are here briefly summarised in three categories.

**Questions**

Q1: What were you thinking about during the interview?

Q2: If you could do the interview again, with exactly the same material, would you do anything differently?

Q3: Are you nervous when doing oral tests? If yes, why? Can you think of anything which the examiner could do to make the experience less stressful? If no, can you imagine any circumstances in which you would feel nervous?

Section 4: **Analysis of Data**

4.1 **Rater responses**

**Question 1**

Rater responses were generally in keeping with the predictions made when the tasks were designed (see pages 234 - 235), confirming that raters were able to see what each task was designed to elicit.
Question 2

The Rasch estimates of difficulty of the three tasks was, from more to less difficult, Task 2 > Task 1 > Task 3 (see page 292). Three of the five raters correctly predicted this sequence. Rater 2 thought that Task 2 was the most difficult but that Task 1 was the easiest, and rater 3 predicted that Task 3 was the most difficult and Task 1 the easiest.

Rater 5 made a most perceptive comment when saying that from merely examining the three tasks one would have assumed that Task 3 would prove to be the most difficult, but after rating the same students doing all three tasks one is forced to reconsider this view.

Two issues must be made raised in regard to the responses to this question. Firstly, the topic chosen for the interview seems to be a major consideration in assessing its difficulty and, from the responses to questionnaires by students was also a major consideration in assessment of the "fairness" of the task. The topic chosen for the task comes up time and time again in most responses, and is clearly an area for much further research. We will return to this in Chapter 11. Secondly, assessment of the difficulty of a task does not seem to be possible on any a-priori basis. This only becomes possible after tasks have been trialled and seen by expert (reliable) judges.
Question 3

When asked which task produced the most assessable language, three of the five judges preferred Task 3 on the grounds of the choice of topic. Two raters preferred Task 1, one because of the choice of the topic. Rater 3, who preferred Task 1 did, however, argue that Task 2 allowed greater discrimination between the very good students and the average students. This is an interesting response in relation to rater 3’s response to question 2 in which Task 3 was isolated as the most difficult: rater 3 does seem to be inconsistent in the way in which task difficulty and discrimination are viewed.

Question 4

Three of the five raters responded that Task 2 was unfair, for the same reasons which were provided for its perceived difficulty in Question 2. Rater 3 thought that none of the tasks were unfair, whilst rater 2 did specify that although Task 2 was difficult, the students should have been able to tackle it much better than the rater thought they did.

Question 5

Although rater 3 would have changed Task 3, without specifying what would need to be changed, this was a minority view. The other four raters all said that they would change Task 2. This assessment of Task 2 on the part of the raters is very much in keeping with the views
of the students expressed in their questionnaire responses.

**Question 6**

No other comments were made on the tasks used in the research, although rater 5 echoed previous views that Task 3 was particularly well designed and useful.

**Question 7**

Responses to this question were virtually unanimous: the rating scale as it stands is viewed as being too long for practical use when one is rating students, as it is difficult to "internalise" or "memorise." It was suggested that the descriptors could be made more concise by producing a list of points that raters must look for.

Rater 1 added that if the same format is to be maintained, then it would aid the rater if the first paragraph in each band descriptor referred to the same factor; for example, hesitation in the first paragraph of the descriptor for each of the bands.

Rater 2 argued that Fluency was a component of "communication," and that it would be useful to rename the rating scale. Rater 4 said in addition that it was difficult to distinguish between bands 3 and 4 as these were very similar. With reference to the ICCs for Fluency in the reliability study (see page 288) we may see whether or not rater 4's intuitions are correct. On Task 1 rater 4 was indeed more likely to move from band 3 to
band 5, but band 4 was confused with band 5 and not band 3. There was no problem at all with the use of these bands on Task 2, while on Task 3 band 4 was used in a distinctively separate way, although for a smaller ability range than either band 3 or 5. It would therefore appear in this case that the retrospective report of rater 4 is not consistent with actual rating patterns.

Rater 5 added that in bands 2 and 3 it is impossible to know why students are pausing, and this caused problems when rating. This perceptive comment relates directly to the discussion on high and low inference criteria for the descriptors (see pages 132 – 135, 204) which was considered when developing the rating scales. It is also important to note that it is with precisely these bands that the highest validity coefficients were discovered (see Table 9.23 in the Appendix to Chapter 9). From rater 5’s own rating pattern (page 288) we may also note that the two bands were adequately distinguished, with the exception of the use of the bands on Task 3 where band 2 was an extreme and therefore not reported.

We may therefore conclude that irrespective of the advantages in research terms of using high inference categories and their potential validity, there will be raters who prefer using rating scales whose band descriptors are termed in much more concrete terms, namely, referring only to strictly observable phenomena in a behavioral sense. It may be that here there is a necessary distinction between rating scales which are
designed for research purposes and rating scales which are designed for actual use in testing programmes.

Question 8

Three raters replied in the negative to this question. Rater 4 did suggest that there were grammatical elements in the Fluency scale, but could not provide specific examples of this. Rater 5 suggested that "confidence in ability to get things right first time" might be more appropriate in the Lexis section of the Accuracy scale, observing that often weak candidates demonstrate confidence even though they do make errors. Although this may be true in a classroom setting (for which no data are available), in the data for this study weak candidates were not observed to display such confidence under test conditions. Further investigation of this possibility could be the object of other studies.

Question 9

Three examples of criteria not represented in the scale were given by raters 4 and 5. These were: humour, the "advanced" ability to create laughter in the interview, the speed of delivery, and flexibility (presumably as contained in the ELTS scale). The issue of speed of delivery was also raised by rater 5 in relation to Question 7. It had been assumed that speed of delivery would be covered by the criterion of "pausing" in the Fluency scale, but rater 5 clearly did not see this as
appropriate. The criteria of humour and flexibility (undefined at present) must be the subject of further research into oral testing.

**Question 10**

Two raters stated that they did not rely on any other criteria than those contained in the Fluency rating scale, while one stated that although no other criteria were used, there was a "gut feeling" regarding the level of the students. Rater 2 wondered whether or not pronunciation could have been included in Fluency, although it was stated that such issues were discounted in actual rating. Only rater 5 said that they were influenced by the notions of "flexibility" and "initiative" as contained in the ELTS scale, as the Fluency scale did not take into account the "passiveness" of the students on some tasks. This comment could very well be associated with the view of rater 2 that "communication" may very well be a superordinate of Fluency, and the observation of rater 1 that harshness in rating was a function of the reluctance of students to take a lead in the discussion. It may very well be that the notion of "communication" as suggested by rater 2 needs a great deal more research and operational definition.
Question 11

Rater 2 argued that the marking range was restrictive, saying that more bands should have been available. The rater stated that particular problems were encountered between bands 3 and 4. This observation can be confirmed from the ICCs (see page 288) of rater 2’s scoring, in which band 4 is either subsumed by band 3 or confounded with bands 3 and 5. However, this is not evidence to suggest that more bands were needed; it is evidence to suggest that Rater 2 did have some difficulty in using the range of bands already available.

Rater 3 said that the middle bands were mainly used, and the extremes were not used. With reference to rater 3’s ICCs this is certainly true of ratings on Task 1, although the top ratings were used on Tasks 2 and 3. Lower bands were used only to a very limited extent, confirming the perceptions of rater 3.

Rater 5 stated that band 1 was not used. This was only true on Task 3, although on the other Tasks it was an extreme. Raters 1 and 4 said that they did use the entire band range, and this is borne out by their ICCs (see page 288).

Question 12

Although this question was omitted for rater 4 it had previously been answered. The majority verdict was that the rating scale needed to be shortened for practical use.
**Question 13**

The feeling that the rating scale was too long for practical use was once again echoed. Rater 2 expanded a little on the title of the scale, saying that if it were called "communication" then flexibility and fluency would be subcomponents of the scale. The rater also suggested, as previously, that there should be a wider band range and, if this were not possible, then the lowest and highest bands should be provided with descriptors.

**Question 14**

Turning to the Accuracy scale, all raters thought that it was more practical and useful than the Fluency scale, although the comments regarding its length and "wordiness" were repeated. The consensus was that the Accuracy scale was easier to use than the Fluency scale, rater 2 stating that there was enough information, clearly spelt out (unlike the ELTS scale), and raters 4 and 5 agreeing that it gave much more concrete information on what to look for when rating a student. Rater 5 alone suggested that the lexis column in the rating scale was still too theoretical, whilst generally approving of the Accuracy scale.

**Question 15**

The answer to this question was unanimously negative, with rater 4 reiterating that the descriptors were too detailed.
Question 16

All raters were reasonably content with the definition of Accuracy used. Rater 4 raised the issue of pronunciation, saying that it was part of the concept of accuracy. The rater was, however, unsure whether the mistakes noticed were indeed mistakes because of familiarity with the pronunciation of Greek speaking students. This would tend to suggest the need to use raters unfamiliar with the accent/pronunciation of the students used in a study if pronunciation is to be considered, and compare their rating patterns with those of raters familiar with the pronunciation of the students.

Question 17

In response to this question the raters unanimously said that they did not rely on criteria external to the scale. Rater 4, however, did say once again that there was a "gut feeling" regarding the ability of the students which comes from teaching students at this level. Rater 4 expanded on this in response to Question 23.

Question 18

Only rater 3 stated that it was impossible to use the entire band range on the Accuracy scale, and this can be confirmed from this rater's ICCs (see page 289). On Tasks 1 and 3 bands 3 and 4 were used to cover the largest range of ability, whilst on Task 2 only band 4
was used to any great extent. Rater 1 did use the entire band range, although band 5 was subsumed under band 4 in Task 3, whilst rater 5 did not use the lower bands on Task 3. The intuition of the raters can therefore be seen to be reasonably accurate.

**Question 19**

Once again, all raters believed that the band descriptors on the Accuracy scale were too long and detailed for practical use in a testing situation. Rater 4 suggested that verb forms could be highlighted as it was believed that this helps to isolate ability more easily than other criteria.

**Question 20**

No raters had any other comments to make.

**Question 21**

Raters 1, 2 and 3 all mentioned the need for a wider band range. It has already been seen in the discussion of question 11 that rater 2 did not distinguish between bands 3, 4 and 5 on the Fluency scale, but had no problems with regard to the Accuracy scale. Rater 1 did not use band 6 on the Fluency scale significantly on Task 3 (see page 288), whilst band 4 was subsumed under band 5 on the Accuracy Scale for Task 2 and band 5 was subsumed under band 4 for Task 3. The view of rater 3 must be treated with much more caution, as it has already been
seen that this rater only used a very limited range of available bands. Rater 1 suggested a band range of 1 - 10 or even 1 - 20. This would seem to be unjustified on the basis of the range of bands which the raters were actually capable of distinguishing even on a scale of 0 - 5 or 0 - 6.

**Question 22**

Raters 2, 4 and 5 all answered that they believed they were fair in their rating of students. Rater 2 mentioned one student who was particularly quiet, who was also mentioned by rater 4 in Question 21. Rater 3 said that her ratings were lenient, which indeed they were on all scales across all tasks (see pages 288 - 290). Rater 1 said that ratings were harsher especially if the student remained silent when asked questions or if they made "basic errors." This is confirmed by the rank order of judge severity (page 279) in which rater 1 is the harshest on Accuracy for all tasks, but the second harshest on Fluency and the third harshest when rating on the ELTS scale.

It would therefore seem that raters do indeed have a very good impression of how lenient or harsh they are in their own rating behaviour, irrespective of whether that rating is reliable or not.
Question 23

Only two raters had any general comments to make. Rater 3 merely echoed previous views that the band descriptors were too detailed, and that Task 3 was rather too difficult. Rater 4 had much more to add.

Rater 4 made five extra comments. Firstly, it was suggested that the presence of recording equipment made both the students and the examiner tenser than they would otherwise be. There may be a case for considering carefully the location and obtrusiveness of any equipment necessary, in order to reduce any test anxiety in those students who may be affected by its presence.

Secondly, rater 4 suggested that the tasks used in the study were not related to the age or the ability level of the students.

Thirdly, it was advocated that there should be a distinction drawn in the Accuracy scale between "basic" and "advanced" grammar. The view expressed seems to be that students may be fluent but still make grave errors in accuracy. It would appear that here is a very clear indication that rater 4 was confounding the two scales, and found it difficult to give, say, a high mark on the Fluency scale and a low mark on the Accuracy scale. This comment on the part of rater 4 may indicate that raters do not feel comfortable with giving radically different ratings to the same student on the same task irrespective of the number of rating scales which they are using. This is closely related to the fourth comment. This rater
commented that although these bands were appropriate for
the ability level of the students actually used in the
study, these students may not be able to pass a written
examination set at the same level, specifying the
University of London GCE as an example. This comment
seems to reveal the need on the part of this rater to
have uniform grades or bands for any given student
irrespective of the skill being tested. This may indicate
a larger need on the part of raters for this kind of
consistency, and hence result in cross-contamination not
only between oral rating scales but also the rating of
different skills, unless the raters are capable of
perceiving differences between those skills and
abilities. This raises a major problem in testing and the
validation of language tests: to what extent do the
scores reflect the (often unexpressed) theoretical views
of the raters themselves? Further, whether raters are
trained (cloned?) or not, is it possible to find raters
without already predetermined views of the nature of oral
proficiency? It would seem to be a requirement of all
future work in the validation of oral rating scales that
the theoretical and/or practical views of the raters used
in the validation study be investigated and described. If
this is not done, it may be the case that the rating
scales themselves have little role to play in the way
raters award scores, and the validation study would be
confounded by constructs which are not accounted for in
the design. An alternative approach would be to use more
naive raters.

The fifth point raised by rater 4 goes further to exemplify this position. In response to Questions 10 and 17 this rater reported the influence of what was termed a "gut feeling" about the ability level of the students being rated. This is here expanded upon, when rater 4 strongly suggests that no teacher should examine students of a level which they have not taught before. This can be taken to mean that the classroom experience of students at a particular ability level can be translated directly into rating accuracy. Any rater then, is bound to bring a great deal of experiential and conceptual baggage to the oral rating process. Rater training - or the absence of rater training - may then not mean a great deal when attempting to do research into the validation of oral language tests. Results are likely to be contaminated from the very beginning by the prior experience which any rater brings to the process of rating.

This may appear to be a rather bleak conclusion for the process of the validation of oral language tests. However, the answer would seem to lie in the number of raters which is employed in research. In a future research study it may be appropriate to attempt to use smaller numbers of students and, perhaps, only two tasks, but a very large number of raters in order to be able to assume that any bias introduced into the rating process from prior experience will constitute random variance. The nature and effect of the conceptual baggage which a
rater brings to the rating process even without training is clearly an area for further research.

4.2 **Student responses**

4.2.1 **Ability level of the sample**

Of the eight students selected for this part of the study, one (student 4 in the transcripts) was excluded from the validity study as anxiety kept her from speaking at a volume which could be adequately recorded for scoring. The ability levels of the other students are given in Table 10.3 in the Appendix to Chapter 10. It can be seen that the selection is from the more able students.

4.2.2 **Summary of findings**

The data collected may be summarised in three categories.

1. Concern of students of the amount of processing time which is required when speaking in a second language needs to be taken into account. It would appear that preparation time and the use of a group discussion task tends to be preferred by students who are aware of this problem.

2. Students seem to engage in dual processing: at the moment of speaking they are concerned with the content of their speech (the "what") and the way in which they communicate their message (the "how"). Planning ahead appears to be concerned with anticipating the next
question in interview situations.

3. Anxiety would appear to be associated with the unnaturalness of the speech event as a test. It is unlike talking that occurs in any other context. In particular, students dislike the gaps in talk that occur and are aware of the fact that the interviewer is assessing them. Students who are natural introverts may need special treatment, and interviewers need to be able to distinguish between ability as defined by a rating scale and character traits which may affect their assessments. Once again, the group discussion task may help to overcome these problems to some degree, and may account for why most students said that they would prefer to do Task 3 to the others if given a choice.

In the interests of reducing anxiety related test method facets, it would also seem appropriate to recommend that raters do not write on notepaper or record scores during a live oral interview unless there is more than one interlocutor. Inhibition caused by the practice of writing whilst engaging in conversation has also been noticed by other researchers (Rodriguez, 1984).
Chapter Eleven

Discussion and Conclusions

Section 1: Preliminaries

Before turning to the main discussion of this chapter, which must inevitably centre around the validity study of Chapter 9 and an assessment of the data-driven approach to scale development, it is necessary to comment upon the interpretability of the results presented in Chapter 9.

Firstly, it has been suggested in this study that validity cannot be properly assessed unless test method facets can be either controlled or accounted for. It is argued that the data presented in Chapter 7 provides sufficient evidence to suggest that affective factors did not significantly effect scores. In particular, it is claimed that test anxiety, although clearly a factor in the oral testing situation, is not such as to cast doubt upon the reliability of the rating procedure.

Secondly, the reliability studies of Chapter 8 show that we may have confidence in the scores obtained and used in the validity study. Reliability estimates were consistently high irrespective of the method of estimation used, with considerable problems occurring with only one rater who was excluded from the validity study as a result. Variation in rating patterns was discovered in both the G-study and the Partial Credit Model study, although in neither case was this variation great enough to significantly decrease reliability.
estimates. It was further discovered that, although there was a significant task effect, this too was not large enough to reduce reliability estimates. In the case of both raters and tasks the outfit statistics from the Rasch Partial Credit study were not significant.

Having taken these test method facets into account in the study, and having controlled for other test method facets such as interviewer, rater training, place of test, variation in elicitation techniques and time, it is claimed that sufficient rigour has been used in the design of the study to have confidence in the results of the investigation into the validity of the rating scales.

Section 2: Validity and the use of the data-driven approach to scale development.

2.1 The lack of divergent validity

In the concurrent validity study tentative evidence was presented to suggest that an oral trait could be isolated, but there was no evidence whatsoever for concurrent divergent or convergent validity of any components of the oral trait, such as fluency or accuracy. Nor was there any suggestion of concurrent validity between the rating scales and teacher assessment of student abilities.

Validity by group differences was confirmed in the Fluency and Accuracy rating scales, but it has been argued that this is not sufficient evidence upon which to make non-trivial claims.
The Multitrait-Multimethod study and the Maximum Likelihood Confirmatory Factor Analysis revealed convergent validity for all rating scales but provided no evidence for divergent validity. The magnitude of the correlations made this conclusion inevitable. Two possible explanations of the data from these studies were suggested:

(a) Each of the rating scales used could tap a single, global trait, rather than separate component traits, or
(b) there is significant cross-contamination in the use of the rating scales.

It can now be suggested that the second of these explanations is the more likely, and would account for the fact that in the ML study the only source of unique variance which could be isolated belonged to the task facet, which was not found to be very large in the reliability study. That this was happening was indicated in the retrospective study in Chapter 10 (see pages 356 - 357), where rater 4 clearly indicates that there is a lack of distinction between the concepts of fluency and accuracy. If this is the case then having to rate a student on a Fluency rating scale and an Accuracy rating scale at the same time would be hypothesized to increase the degree of cross-contamination. It is at this point that the study by Yorozuya and Oller (1980), despite other problems with the design of their study, is
extremely helpful. They argue (ibid., 145) that:

"...marks assigned by raters may be influenced on each scale by the rating(s) they have just assigned on the previous scale(s) and by the general impression they are forming in the process of assigning the marks. If this were so, and if raters did tend to assign similar marks across the four scales rated on the same occasion, we would find stronger correlations across scales rating on the same occasion than when they are rated on separate occasions."

A similar point is made by Hughes (1981: 179). Yorozuya and Oller called this phenomenon a "halo effect" which they claimed accounted for approximately 10% of the variance in their study (i.e., not attributable to a g-factor). However, in the light of this study it would also seem possible to turn their conclusion on its head, and say that 90% of variance was due to cross-contamination and 10% due to other factors! Yorozuya and Oller were arguing, after all, that the small amount of significant variance which they had detected in their tests did not represent an oral trait but error variance, in defence of the Unitary Competence Hypothesis. They therefore claimed that this variance was "an artefact of the method of rating multiple scales on a single hearing of each interview" (ibid., 146). The MTMM and ML studies in this research strongly suggest that raters are incapable of maintaining a conceptual distinction between component rating scales in the process of rating if the ratings are done at the same time, as they were in this
study (see page 267). It may be suggested that, if the raters are using two or three rating scales, they should do the ratings from recordings on two or three separate occasions. However, this would not only seem impractical in a real testing situation, but would probably also introduce the test method facet of time which would have to be estimated. If the retrospective comments of rater 4 are also generalizable to other raters, it may even then not lead to evidence of discriminant validity.

It is therefore argued here that the apparent lack of discriminant validity in the study of the rating scales which were developed using a data-driven approach is the result of cross-contamination. That this is indeed the case is strongly supported by the concurrent study between the components of the FCE rating scale and the Accuracy, Fluency and ELTS rating scales. The rotated Principal Axis Factor Analysis results in Table 9.6 in the Appendix to Chapter 9 clearly shows that the results are generated by cross-contamination in both tests, because ratings are made at a particular time on all components of the rating scale.

Given the current status of our knowledge about the rating process and the structure of oral proficiency, it would therefore seem appropriate to recommend the use of a single rating scale in oral assessment for the purposes of live testing. The development of componential rating scales in tandem with investigating ways of trying to overcome the problem of cross-contamination should, for
the moment, remain in the realm of research.

In particular, the finding of very strong cross-contamination in this research does not help to throw any light on the theoretical distinction which would predict the difference between the development of accuracy and fluency (Hatch, 1974; Seliger, 1980; Brumfit 1984; Clahsen, 1985; Bialystok and Sharwood Smith, 1985; Ellis, 1990a). Thus, when Skehan (1987: 206) recommends the use of convergent-discriminant techniques designed to validate oral tests to examine the epistemological status of these theoretical but largely untested assumptions, it should be added that these techniques will not always work adequately because of confounding factors.

Ways of researching and overcoming these problems will be discussed below in section 4, directions for future research.

2.2 Validity in the Partial Credit Model study

2.2.1 A methodological note

In assessing the Partial Credit Model study of validity in Chapter 9 we must first begin with a note on methodology. It was stated on page 328 that the 0 and 1 points on the validity formula developed in this study to estimate the validity of individual bands were absolutes, but that the shape of the curve and the points along it were relative to the method of calculation. The interpretation of the validity coefficients are therefore, like all reliability and validity
coefficients, relative. That is, a judgement has to be made by the researcher regarding the magnitude of any figure which is to be termed "acceptable" or "unacceptable".

In Partial Credit Models it is believed that there is a way to objectify these decisions. This method would involve assessing the variation of $\delta_{ik}$ across tasks (that is, the variation seen to occur in the columns of Table 9.22 in the Appendix to Chapter 9) in terms of whether the amount of variation recorded would be expected to occur by chance. This would initially involve calculating the log likelihood of the $\delta_{ik}$ statistic for two models: (i) tasks separate and (ii) tasks combined. The maximum log likelihood could then be calculated as:

$$-2 \log \frac{L_1}{L_2}$$

where $L_1$ is the first model and $L_2$ the second model. The figure produced by this calculation would then be checked against the $X^2$ distribution, the degrees of freedom being calculated as $df_{L_1} - df_{L_2}$. If the maximum log likelihood were found to be large and significant, and larger than the number of parameters in the model, the differences would represent real variation between the $\delta_{ik}$ statistics for the tasks. Once the size of significant differences were known, absolute cut-off points could be applied to the validity curve in Figure 9.3 on page 330.
Unfortunately, although FACETS estimates the maximum log likelihood in the analysis of multi-faceted designs, it does not print it (Linacre and Wright, 1990: 15). This is a major disadvantage of this particular software. Until programs which conduct multi-faceted Partial Credit analysis which do print this statistic are available, it is unlikely that further research into the absolute cut-off points in the type of validity coefficients developed for this study can be conducted. However, as soon as such tools are available, there is a great deal of potential for extending the use and power of the new validity statistic presented on page 327.

2.2.2 Coherence and Continuum Validity

In the Partial Credit Model study, two criteria for validity were put forward. The first criterion was that the validity coefficients in Table 9.23 in the Appendix to Chapter 9 should be high in columns, showing that the test method facet did not affect the operation of the rating scale. It is argued that it is a basic principle in scientific enquiry that the method of data collection does not influence the results obtained. We will now call this the "Coherence" criterion, implying that any construct which cannot adequately be measured using different methods is not in itself coherent. The second criterion was that the validity coefficients in Table 9.23 should be high in rows, and that Item Characteristic Curves should be roughly equally spaced across the entire
ability range measured by the rating scale, showing that the band descriptors of the rating scale represent a proficiency continuum. We now call this the "Continuum" criterion.

The only rating scale which meets both of these criteria is the Fluency rating scale, bands 2 to 4, which are given on pages 161 - 162.

From Figure 4.1 on page 152, it can be seen that in bands 2 and 3, the predictions from the discriminant analysis on the basis of the definition of the construct of fluency were highly accurate. In the case of band 4, only one student would have been misplaced in a band 3. This must now be compared with the finding (page 288) that on the ICCs for tasks 2 and 3, band 4 is seen to be slightly narrower than the other bands on the rating scale, meaning that it needed to be rather more general in nature in order to cover a slightly wider ability range. That is, it did not entirely meet the continuum criterion of validity as described above.

It can be claimed with some confidence that the misplacement of the one student in the initial discriminant analysis when the band was being developed is directly related to, and interpretable in the light of, the results of the validity study. Where the predictions from the initial discriminant analysis were completely accurate, the validity coefficients meet both the coherence and continuum criteria. Where one student was misplaced by one band, the band in question does not
meet the continuum criterion in full, but does meet the coherence criterion. This is a clear indication that bands 2 and 3 on the Fluency rating scale are valid, and that band 4 needs some adjustment in the direction of generality and trialling for a second time.

This situation may be compared with that of the Accuracy rating scale. It was shown in Figure 5.1, and in the discussion on pages 201 - 204, that the initial predictions from the discriminant analysis were not accurate, even though the band descriptors did follow current theories in the development of grammatical accuracy from the second language acquisition literature. The validity estimates in Table 9.23 in the Appendix to Chapter 9 and the associated ICCs do not reflect coherence or continuum validity.

It was suggested on page 203 that the explanation for the results could have been that in oral tests raters themselves concentrate more on aspects of ability which are very close to what we have described as fluency, rather than to accuracy. It was further suggested that if this were the case, if the assessment of accuracy were essentially indeterminate, then low validity coefficients would be expected in the validity study. It would be rash to conclude that this is in fact the case, as no evidence can be presented for the validity of the Accuracy rating scale at all. However, it is one possibility that may deserve further research in the future.

Finally, we note that the ELTS global rating scale,
which is an a-priori rating scale included in this study for comparison with the two rating scales developed in the data-driven approach, has the worst validity coefficients of the three scales.

2.2.3 Conclusions

Two important conclusions are drawn from the above discussion. Firstly, the data-driven approach to rating scale development has shown much promise, despite only three bands from one rating scale having been found valid. If the results of the initial discriminant analysis are excellent the researcher would seem to be able to expect adequate validation figures in terms of what we have termed coherence and continuum. It would appear from this study that as soon as there is the slightest error in prediction from the description of bands to the rank order of the sample of students, validation does not seem to be possible.

Secondly, it would seem that the concern expressed in Chapter 4 over the "high-inference" nature of the band descriptors was unfounded. It would seem that the descriptors for three of the bands are coherent and represent a proficiency continuum as we currently understand it. It is because of this result that the claim to validity is not as trivial as it would have been if, for example, bands 2 - 4 of the accuracy scale had been found to be valid instead. Rather, we may genuinely say that (at least part of) the concept of fluency has
been explained and the explanation found to be satisfactory in operational terms. It is therefore possible to generalize from these three bands, and say that students who fall into them are likely to be able to perform in a similar way on other tasks.

This research is therefore significant in that a methodology in the validation of oral rating scales based on a data-driven approach has been attempted, and consistencies observed between the initial analyses and band descriptor construction and the validation study. In the validation study itself, a new statistic has been successfully developed and used which will allow the investigation of individual bands rather than entire rating scales and, using ICCs, provide evidence upon which we may judge why a band descriptor is or is not working adequately. Three band descriptors have been found to meet the validity criteria of coherence and continuum.

Positive evidence of the effect of cross-contamination has been isolated in both the FCE rating process as well as the rating process within this study. It is suggested that this is common to the rating process in oral testing, and ways to overcome this must be found if componential scoring is to be used in live testing.
Section 3: **Assumptions of this research**

It has been assumed throughout this study that it is possible to work within what may be called a "stable trait" paradigm. That is, it has been assumed that, although interlanguage may be continually developing it is not changing radically from minute to minute or second to second. If the latter were the case, then any trait which any test aims to tap validly would be in constant flux and there would be no empirical way to separate trait from method, and no way to claim that any results from any investigation can be said to be related to the trait in question at all. For testing purposes, it would appear necessary to assume that the trait is stable at least for the period of time in which the test is conducted so that a "snap shot" of the trait may be taken. The existence and nature of interlanguage, the degree of stability and the speed at which it changes are current issues for debate within the second language acquisition literature (Gregg, 1990; Ellis, 1990b; Tarone, 1990).

Secondly, the stable trait assumption also applies to the nature of the task. It has been argued with Bachman and Savignon (1986) that the trait must be separable from test method facets if validation is to be possible, and task is one of those facets. It follows that the principle of validation by coherence discussed above requires that the measurement of the trait be stable across tasks or test method facets. If there is
variability of trait across tasks then the trait cannot be isolated or measured. Tarone (1983: 147 - 152; 1985), on the other hand, argues that "capability" underlies performance, that this is heterogenous although internally consistent and describable, but varies by speech style which is related to the nature of the task being undertaken by the student.

This is supported by Douglas and Selinker (1985) and Ellis (1985b) who propose that there are several interlanguages, several capabilities, which depend on the "discourse domain" of the task. This would imply that language tests must be devised to tap every single identifiable discourse domain and every single identifiable style on a continuum.

Skehan (1987: 200) accepts the implications of this for testing, arguing that the problem for language testers working within this paradigm is one of sampling. In designing the studies reported in this project, these implications have not been accepted.

What has been accepted is that there is task variability, but that this is not great enough to dispense with the concept of a stable trait (or part of the underlying competence) which is not entirely task dependent. The finding that task 3 in this study is different from the others, but not so different as to generate large outfit statistics (see page 292) supports this view. It is also accepted that the trait or competence changes over time, resulting in different
performance on tasks at different stages of interlanguage development (as is assumed in Pienemann et. al., 1985). However, with Gregg (1990) it is assumed that synchronic variability is a feature of performance and not of the trait itself which generates that performance. We concur with Gregg (ibid., 370) that:

"The distinction between competence and performance...has pretty much established itself in linguistic theory and (to a lesser extent perhaps) in acquisition research. And rightly so; making the distinction is simply a fundamental prerequisite to progress in the scientific study of language acquisition. There is thus a great deal at stake in any attempt to deny or to blur over the distinction..."

Ellis (1990b) argues that the variability is motivated by the need to examine the "use" of the language for educational purposes, rather than the concerns of linguistics per se. And indeed, variability due to task or style is a legitimate area for research. However, for the purposes of oral language testing, it has been argued that it is essential to assume that the trait is stable and generates the performance. The argument quite simply is that only by making this assumption can one make non-trivial statements about why students perform in the way they perform and, most importantly, generalize from the testing situation to other situations external to it. This has been the prime objective of testing, as argued by Skehan (1988: 211) and discussed on page 9. To accept that tests are needed for
every discourse domain and every speech style would necessarily bring with it the acceptance that no traits may be identified and no validation studies can be carried out. Above all, no results would be capable of generalization, and it is only statements about language learning and testing which can be generalized that are of interest.

The theoretical assumptions underlying this present study have been briefly considered here so that the reader may assess the extent to which the results warrant the conclusions drawn. It also points up the fact that no empirical study can be conducted without theoretical assumptions being made!

Section 4: Suggestions for future research
4.1 Validation of component traits

Divergent validity of the two traits of accuracy and fluency could not be established in this study, mainly because of the effect of cross-contamination. In order to investigate the structure of the oral trait and introduce the possibility of achieving divergent validity, the effect of cross-contamination must be catered for in any future research project.

It is suggested that one way of overcoming this problem is to write rating scale specifications, which would contain as full a description of the trait to be measured as possible and an indication of the number of bands to be used in the rating scale. Two rating scales
would then be produced for each trait by separate scale writers on the basis of the specifications.

Two sets of raters would then be asked to rate a small number of students doing the same task from video on two of the rating scales on different occasions. This would, of course, introduce a time facet. To estimate the effect of this, it would be recommended that at time 1, one set of raters marked the students on rating scale 1 for trait 1, and simultaneously the second group of raters marked the students on rating scale 1 for trait 2. At time 2, the first set of raters would rate the students on rating scale 1 for trait 2, and the second set of raters would rate the students on rating scale 2 for trait 1. The two rating sessions would have to be held as close to each other as reasonably possible.

A Partial Credit model could be used to assess the rating pattern of the raters on the two scales, and an MTMM study could be devised in which the methods were rating scales. It would be hypothesized that rating scales of the same trait would exhibit divergent validity, as the design of the data collection technique would presumably reduce the effect of cross-contamination.

Another approach to this problem may be to use recently developed SOPIs in conjunction with a rating procedure which did not require raters to make any judgements at all. For example, the taped responses of students to a completely standardized elicitation
technique could be transcribed and rated on the basis of purely linguistic criteria hypothesized to be related to two or more distinct traits. This method would not be as powerful as that described above, as it would of necessity have to concentrate only on verbal elements of speech, but it would at least give some insight into whether or not the simple counts of the occurrences of the verbal elements said to be exponents of the separate traits would result in the students being awarded different scores on those traits. This type of design is, in fact, very much like that used by Pienemann et al. (1985) in the construction of their oral observation forms.

Yet another approach to the problem would be to conduct an introspective study. In this design a large number of raters would be used to rate a small number of students taking two or three oral tests on a number of component rating scales. However, prior to doing this, all raters would be interviewed about their conception of the nature of the traits on which they will be asked to rate. These interviews would have to be recorded, transcribed, and the concepts of each rater thoroughly documented. Immediately after this the raters would then be asked to rate the students, and the rating patterns in practice could then be related directly to the introspective data in order to see to what extent prior views of raters affected their use of the rating scale. It may also be possible to arrange for the raters to
introspect as they are rating, commenting on their interpretation of the rating scale, and a retrospective study in which raters are shown their actual rating pattern and ask them to compare this to what they thought they were doing at the time. Such a study would throw a great deal of light onto the rating process which at the present time we know very little about, although if the comments of rater 4 (see pages 353, 356 - 359) are generalizable, the conceptual baggage which raters bring to the rating process would seem to be an area of great concern for oral testing.

4.2 Task and topic

Although it is known that there is a task effect in oral testing, this does not seem to be as great as has often been assumed (see page 292). Nevertheless, we do not know a great deal about task difficulty. Much more research is needed into task difficulty along the lines of Kenyon and Stansfield (1992), but using scores from actual oral tests rather than relying on the judgements of teachers. Further research in this is important, as we have seen that difficulty level does relate to test anxiety (see page 259), and if it is possible for tasks to be scaled on a difficulty continuum, appropriate tasks may be selected for students of certain predicted ability levels in order to control test anxiety.

Similarly, research is needed into the type of language and discourse produced by certain tasks. We have
seen in this study that task 3, the group discussion, was seen by raters as providing the most assessable language, that students perceived it as less stressful and the situation as one in which more "natural" conversation could occur. From the outfit statistics (Tables 8.38 - 8.40 in the Appendix to Chapter 8) task 3 was seen to be scalable on the same dimension as the other tasks, and yet did appear to be operating in a different way from the others to some extent. It would be interesting to see research similar to that of Perrett (1990) into the language and discourse produced in the interview conducted for this type of task.

Finally, no research has been discovered which takes into account task and topic and any possible interaction effects between the two as test method facets. In this study, each of the tasks had its own topic, and so the two could not be studied separately. Yet, in the data from the retrospective study in Chapter 10, the issue of familiarity with the topic was consistently raised, and in the questionnaire data discussed in Chapter 7 the choice of topic was seen to be related to student perceptions of "fairness".

We suggest a research project in which one task is designed with one standardized elicitation procedure, but with two or more different topics. A sample of students would be asked to take the two or more tests which would then differ only in terms of topic, and each of the students rated on a single rating scale. It would be
hypothesized that any difference in the scores would thus be due to the effect of topic. Prior to carrying out this study, it would be recommended that each of the students be interviewed about their background interests, learning history, and any other factors which might influence their ability to talk about one of the topics chosen for the test. The results of the tests could then be related directly to the background of each student in the sample, making it possible to isolate the nature of any topic effect in oral testing.

4.3 Specificity and length of rating scales

Writers from the time of Lado (1961) to Baker (1989), Hughes (1989) and Matthews (1990) have argued that a rating scale must be as specific and detailed as possible. On the other hand, Porter (1991) argues that they should be short and as simple as possible, to be judged on the basis of what they leave out rather than what they include. In the retrospective study, the raters used in this research unanimously expressed the view that the rating scales of Accuracy and Fluency were too long and detailed. Yet, no research has been conducted on the optimum length of rating scales for practical use; the views expressed regarding length are merely the opinions of the writers in the light of their own personal experiences of rating.

Research is therefore needed into the length of rating scales. It is suggested that this could be done by
taking one extremely detailed rating scale and producing from it a series of other rating scales along a cline of decreasing specificity by editing out parts of the scale, until it became just a title with a list of band numbers. Teams of raters would then rate the same students doing the same task on one of the scales independently. Reliability coefficients for each of the teams, and the rating patterns could be compared using a Partial Credit model. It would be assumed that somewhere along the cline the evidence would suggest that optimum reliability had been achieved, and thus empirical evidence would be available to suggest that a certain degree of specificity was preferable to others.

4.4 Validation methodology and tools

It has been argued that the new validation coefficient which relates to bands within scales rather than whole rating scales (page 327) is a large step forward in validation procedures. Validation by coherence and continuum as defined on pages 368 - 371 becomes possible. However, it has also been noted (pages 366 - 371) that there is still a problem with the relative nature of the interpretation of the validity coefficients when they are distant from the extreme points of the validation curve.

Overcoming this problem in the way suggested depends on the production and availability of programs for many-faceted models which provide the appropriate statistics.
for absolute cut-off points to be drawn on the validity curve. When such tools become available, it would be advisable for more research to be done on the new validity statistic developed in this study. The data from this study could be analyzed again using the new programs, or data from other studies, in order to discover precisely how the validity coefficient works, and the magnitude of the difference between coefficients in columns which would indicate a real difference in validity.

4.5 Research into other rating criteria or traits

This study has only dealt with the two traits of accuracy and fluency for reasons stated in the rationale (pages 114 - 118). However, it is not assumed that these are the only component traits which go to make up oral ability. Nor is it assumed that the criteria used in the rating scales to define these traits are the only ones which could be used. Research is needed into the validation of other possible traits and the criteria used to define them.

In the retrospective study, raters mentioned the possibility of measuring ability to communicate, flexibility and the ability to take initiative in the conversation, for example. One rater suggested that the ability to be humorous may be a criterion which could be included within the definition of fluency.
We currently know so little about the structure of oral ability that this area of research is wide open.

Section 5: Conclusions

It may appear, especially from section 4 (above) on areas for future research, that this study has taken the reader from a state of confusion to a higher state of confusion. There is undoubtedly some truth in this. However, in this study the territory of research and practice in oral testing has been staked out carefully and some new problems, both theoretical and practical, revealed and described. Above all, it is believed the data-driven approach to rating scale construction, the validation of bands rather than rating scales and the concepts of coherence and continuum, will provide new insights and contribute to methodological development in future research into oral testing.
References


Alderson, J. C., Krahneke, K. J. and Stansfield, C. W. (Eds) 1987 Reviews of English Language Proficiency Tests. TESOL.


APA 1985 Standards for Educational and Psychological Testing.


Bialystok, E. and Sharwood Smith, M. 1985 "Interlanguage is not a state of mind: an evaluation of the construct for second language acquisition." Applied Linguistics 6, 101 - 117.


Candlin, C. and Murphy, D. F. 1987 Language Learning Tasks. Lancaster Practical Papers in English Language Education; Prentice Hall.


Ellis, R. 1987a *Second Language Acquisition in Context.* Prentice Hall.


Heaton, J. B. 1975 *Writing English Language Tests.* Longman.


Higgs, T. V. (Ed) 1984 *Teaching for Proficiency, the Organizing Principle.* Lincolnwood, IL: National Textbook Company.


396


Ingram, D. 1982 Introduction to the ASLPR. Mimeo.

Ingram, D. 1985a "Assessment and the ASLPR." Paper to the conference of the Curriculum Area Committee Languages, Senior Secondary Assessment Board of South Australia, Adelaide.


James, C. J. (Ed) 1986 Foreign Language Proficiency in the Classroom and Beyond. Lincolnwood, IL.: National Textbook Company.


Larsen-Freeman, D. 1975b "The acquisition of grammatical morphemes by adult ESL students." TESOL Quarterly 9, 409 - 430.


Madsen H. S. and Murray, N. 1984 *Retrospective evaluation of testing in ESL content and skills courses*. Brigham Young University, unpublished manuscript.


Matthews, M. 1990 "The measurement of productive skills: doubts concerning the assessment criteria of certain public examinations." English Language Teaching Journal 44, 2, 117 - 121.

McNamara, T. F. 1990 "Item Response Theory and the validation of an ESP test for health professionals." Language Testing 7, 1, 52 - 76.


Quinones, J. (no date) "Independent rating in oral proficiency interviews." Central Intelligence Agency.


Sunderland, J., Yixing, Z., and Barr, B. 1988 "To be an explainer: a reporting test of oral proficiency." English Language Teaching Journal 42, 1, 28 - 33.


409


Wright, T. 1987 "Instructional Task and Discoursal Outcome in the L2 Classroom." In Candlin C. and Murphy, D. F. (Eds) 1987 Language Learning Tasks. Lancaster Practical Papers in English Language Education; Prentice Hall, 47 - 68.


Appendix to Chapter Two

Table 2.1. Correlation coefficients of the ratings of two judges for the FSI global and component scores (Clark and Swinton, 1980)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>.79</td>
<td>.63</td>
<td>.82</td>
<td>.70</td>
<td>.72</td>
<td>.82</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>.58</td>
<td>.59</td>
<td>.57</td>
<td>.46</td>
<td>.52</td>
<td>.54</td>
</tr>
<tr>
<td>Grammar</td>
<td>.77</td>
<td>.58</td>
<td>.80</td>
<td>.67</td>
<td>.66</td>
<td>.79</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>.72</td>
<td>.52</td>
<td>.76</td>
<td>.64</td>
<td>.66</td>
<td>.75</td>
</tr>
<tr>
<td>Fluency</td>
<td>.70</td>
<td>.55</td>
<td>.71</td>
<td>.62</td>
<td>.65</td>
<td>.75</td>
</tr>
<tr>
<td>Comprehension</td>
<td>.70</td>
<td>.52</td>
<td>.73</td>
<td>.63</td>
<td>.67</td>
<td>.76</td>
</tr>
</tbody>
</table>

Table 2.2. Correlation coefficients of the ratings of two groups of raters for the FSI global and component scores (Hendricks et. al., 1980).

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Global</th>
<th>Pron.</th>
<th>Grammar</th>
<th>Vocab.</th>
<th>Fluency</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pronunciation</td>
<td>.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grammar</td>
<td></td>
<td>.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td></td>
<td></td>
<td>.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluency</td>
<td></td>
<td></td>
<td></td>
<td>.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comprehension</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.89</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3. The relationship between ACTFL scores and years of study (Magnan, 1986a)

<table>
<thead>
<tr>
<th>Year</th>
<th>Range</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st. Year</td>
<td>N-mid - I-mid/high</td>
<td>I-low/mid</td>
</tr>
<tr>
<td>2nd. Year</td>
<td>I-low - Advanced</td>
<td>I-mid</td>
</tr>
<tr>
<td>3rd. Year</td>
<td>I-mid/high - Adv/Adv+</td>
<td>I-high/Adv</td>
</tr>
<tr>
<td>4th. Year</td>
<td>I-mid - Adv+</td>
<td>Advanced</td>
</tr>
</tbody>
</table>

412
Table 2.4. FSI rating components reported to discriminate at each level in the FSI rating scale (Adams, 1980).

<table>
<thead>
<tr>
<th>FSI level</th>
<th>Discriminating factors in descending order of significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0+ - 1</td>
<td>Vocabulary</td>
</tr>
<tr>
<td>1 - 1+</td>
<td>Fluency, Comprehension, Grammar, Vocabulary</td>
</tr>
<tr>
<td>1+ - 2</td>
<td>Comprehension, Grammar, Accent, Fluency</td>
</tr>
<tr>
<td>2 - 2+</td>
<td>Fluency, Comprehension, Accent, Vocabulary</td>
</tr>
<tr>
<td>2+ - 3</td>
<td>Grammar, Accent, Vocabulary, Comprehension</td>
</tr>
<tr>
<td>3 - 3+</td>
<td>Comprehension, Fluency, Grammar</td>
</tr>
<tr>
<td>3+ - 4</td>
<td>Vocabulary, Accent, Grammar</td>
</tr>
<tr>
<td>4 - 4+</td>
<td>Grammar, Vocabulary</td>
</tr>
</tbody>
</table>

Table 2.5. Correlations between the TSE component rating scales and the FSI Oral Proficiency Interview (Clark and Swinton, 1980: 19).

<table>
<thead>
<tr>
<th>TSE</th>
<th>FSI Total Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronunciation</td>
<td>.77</td>
</tr>
<tr>
<td>Grammar</td>
<td>.73</td>
</tr>
<tr>
<td>Fluency</td>
<td>.76</td>
</tr>
<tr>
<td>Comprehension</td>
<td>.76</td>
</tr>
</tbody>
</table>
Table 2.6. Correlations between the SOPIs and the OPI
(Stansfield, 1990a: 231)

<table>
<thead>
<tr>
<th>Language</th>
<th>Same Rater</th>
<th>Separate Raters</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>.96</td>
<td>.90</td>
<td>.93</td>
</tr>
<tr>
<td>Portuguese</td>
<td>.93</td>
<td>.93</td>
<td>.93</td>
</tr>
<tr>
<td>Hebrew (USA Version)</td>
<td>.93</td>
<td>.96</td>
<td>.92</td>
</tr>
<tr>
<td>Hebrew (Israeli Version)</td>
<td>.95</td>
<td>.94</td>
<td>.93</td>
</tr>
<tr>
<td>Indonesian</td>
<td>.98</td>
<td>.94</td>
<td>.93</td>
</tr>
<tr>
<td>Hausa</td>
<td>.91</td>
<td>.8</td>
<td>.84</td>
</tr>
</tbody>
</table>

Table 2.7. A Multitrait-Multimethod Validation matrix with skills as trait and rater/item type as method: M/C = Multiple choice, O/E = open ended. (Dandonoli and Henning, 1990).

<table>
<thead>
<tr>
<th>Skill</th>
<th>Speaking</th>
<th>Writing</th>
<th>Listening</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RatA RatB</td>
<td>RatA RatB</td>
<td>M/C O/E</td>
<td>M/C O/E</td>
</tr>
<tr>
<td>RA</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaking</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RB</td>
<td>.97</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RA</td>
<td>.85</td>
<td>.86</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Writing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RB</td>
<td>.92</td>
<td>.88</td>
<td>.87</td>
<td>1.00</td>
</tr>
<tr>
<td>M/C</td>
<td>.80</td>
<td>.79</td>
<td>.77</td>
<td>.80</td>
</tr>
<tr>
<td>Listening</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O/E</td>
<td>.76</td>
<td>.71</td>
<td>.74</td>
<td>.77</td>
</tr>
<tr>
<td>M/C</td>
<td>.83</td>
<td>.79</td>
<td>.80</td>
<td>.82</td>
</tr>
<tr>
<td>Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O/E</td>
<td>.76</td>
<td>.75</td>
<td>.78</td>
<td>.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.86</td>
<td>.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 2.8. Means and Standard Deviations of Estimated Person abilities on a single test of speaking associated with ACTFL oral proficiency levels (Dandonoli and Henning, 1990).

<table>
<thead>
<tr>
<th>Proficiency</th>
<th>N</th>
<th>Mean</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice Low</td>
<td>1</td>
<td>n.a</td>
<td>n.a</td>
</tr>
<tr>
<td>Novice Mid</td>
<td>6</td>
<td>-3.50</td>
<td>0.51</td>
</tr>
<tr>
<td>Novice High</td>
<td>2</td>
<td>-1.25</td>
<td>0.41</td>
</tr>
<tr>
<td>Intermediate Low</td>
<td>24</td>
<td>-2.16</td>
<td>0.35</td>
</tr>
<tr>
<td>Intermediate Mid</td>
<td>27</td>
<td>-0.70</td>
<td>0.16</td>
</tr>
<tr>
<td>Intermediate High</td>
<td>29</td>
<td>-0.05</td>
<td>0.13</td>
</tr>
<tr>
<td>Advanced</td>
<td>14</td>
<td>1.36</td>
<td>0.18</td>
</tr>
<tr>
<td>Advanced Plus</td>
<td>8</td>
<td>2.32</td>
<td>0.29</td>
</tr>
<tr>
<td>Superior</td>
<td>7</td>
<td>3.97</td>
<td>0.53</td>
</tr>
</tbody>
</table>
Appendix to Chapter Four

Table 4.1. Frequency distribution of scores on the ELTS test.

<table>
<thead>
<tr>
<th>Band</th>
<th>Scores</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band 9</td>
<td>I</td>
<td>(1)</td>
</tr>
<tr>
<td>Band 8</td>
<td>I</td>
<td>(1)</td>
</tr>
<tr>
<td>Band 7</td>
<td>III</td>
<td>(4)</td>
</tr>
<tr>
<td>Band 6</td>
<td>I</td>
<td>(8)</td>
</tr>
<tr>
<td>Band 5</td>
<td>III</td>
<td>(4)</td>
</tr>
<tr>
<td>Band 4</td>
<td>III</td>
<td>(3)</td>
</tr>
</tbody>
</table>

Table 4.2. Multivariate test results for prediction from eight explanatory categories of Fluency to ELTS scores.

<table>
<thead>
<tr>
<th>Categorial Variable</th>
<th>Wilks' Lambda</th>
<th>F</th>
<th>df</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band Scores</td>
<td>0.02</td>
<td>2.06</td>
<td>32, 34</td>
<td>0.02</td>
</tr>
<tr>
<td>1st Discriminant Function</td>
<td>X² = 52.903</td>
<td>p = .00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd Discriminant Function</td>
<td>X² = 27.704</td>
<td>p = .14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3. Spearman Rho for individual explanatory categories of Fluency and ELTS scores.

<table>
<thead>
<tr>
<th>Category</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELTS Scores</td>
<td>-.23</td>
<td>.40</td>
<td>-.60</td>
<td>.63</td>
<td>-.38</td>
<td>-.17</td>
<td>.15</td>
<td>-.83</td>
</tr>
</tbody>
</table>

Table 4.4. Means for students scoring different bands for explanatory category 1 in the definition of fluency.

<table>
<thead>
<tr>
<th>Bands</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.67</td>
<td>0.50</td>
<td>0.63</td>
<td>0.50</td>
<td>4.50</td>
</tr>
</tbody>
</table>

Table 4.5. Means for students scoring different bands for explanatory category 2 in the definition of fluency.

<table>
<thead>
<tr>
<th>Bands</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>12.33</td>
<td>19.25</td>
<td>28.38</td>
<td>39.50</td>
<td>13.00</td>
</tr>
</tbody>
</table>

Table 4.6. Means for students scoring different bands for explanatory category 3 in the definition of fluency.

<table>
<thead>
<tr>
<th>Bands</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>16.67</td>
<td>15.50</td>
<td>8.75</td>
<td>6.00</td>
<td>2.00</td>
</tr>
</tbody>
</table>

415
Table 4.7. Means for students scoring different bands for explanatory category 4 in the definition of fluency.

<table>
<thead>
<tr>
<th>Bands</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.67</td>
<td>2.00</td>
<td>5.38</td>
<td>8.75</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 4.8. Means for students scoring different bands for explanatory category 5 in the definition of fluency.

<table>
<thead>
<tr>
<th>Bands</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5.33</td>
<td>5.75</td>
<td>6.00</td>
<td>3.25</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 4.9. Means for students scoring different bands for explanatory category 6 in the definition of fluency.

<table>
<thead>
<tr>
<th>Bands</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>9.00</td>
<td>12.75</td>
<td>12.00</td>
<td>9.25</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Table 4.10. Means for students scoring different bands for explanatory category 7 in the definition of fluency.

<table>
<thead>
<tr>
<th>Bands</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.67</td>
<td>2.25</td>
<td>2.50</td>
<td>3.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 4.11. Means for students scoring different bands for explanatory category 8 in the definition of fluency.

<table>
<thead>
<tr>
<th>Bands</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5.67</td>
<td>5.75</td>
<td>0.50</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>


In the following transcript, the categories of Fluency listed on pages 135 - 149 are given between slashes. The notation "/3/" would therefore indicate grammatical planning hesitation. "A>" indicates the speech of the interviewer and "B>" the speech of the student.

Utterances enclosed by square brackets indicates speech overlap, while round brackets indicate noises, the removal of identifying information or sections of
inaudible speech.

The oral interview in which this sample was collected consisted of an introductory phase in which the student was put at his/her ease through the use of general questions about home or work, followed by questions on a text which had been read previously. After this, students are asked about their future plans, followed by a brief wind-up, mainly involving leave taking.

When reading the following transcript it must be remembered that the researcher was present at the oral interview, and when coding the transcript was working with audio recordings of the interview data.

Interview 8: ELTS Medical Module: Band 7
A> Hello:: I'm (NAME) B> I'm (NAME) not (NAME) A> (NAME) [yes B> yeah] A> that's what that's what I've got that's [what B> ah] A> that's what I said [the B> but] they called me (NAME) A> ah yeah they it was the the person who I asked him to call you but er never mind: anyway I understand you're a pharmacist B> yeah A> is that right tell me something about your work B> ah: erm /2/ I work at I don't know if you know Cyprus at all A> yes I do B> I work at (NAME) hospital A> hm m B> and I work as a pharmacist as a hospital pharmacist we are three pharmacists there we actually erm /2/ dispense medicines and er: /4/ of course we have other things like preparing: some kinds of medicines the (inaudible) A>
yeah B> sceptic solution: A> hm: so it’s generally
dealing with drugs in the hospital that B> yeah and of
course we supply the medicines to most other wards and
casualty and we’re actually responsible: A> yeah so you
have to check to make sure that none of them go missing
B> yeah but their expiry dates erm: /4/ if they’re kept
well if they’re kept properly A> yeah it sounds very
interesting is (NAME) a big hospital B> no actually it
it’s a small one A> hm: right but you you live out there
in (NAME) B> I have to (laughter) A> yeah B> because I am
er: I am /2/ on call every other day A> yeah B> er but I
I /4/ come from Nicosia A> hm so you would you like to
transfer to Nicosia General if [you B> yeah] were given
half [the chance B> for sure] A> or Makarios probably
even nicer B> yes [(laughter)] A> yeah because Makarios
is a modern hospital whereas Nicosia is beginning to fall
to bits B> yes A> a little isn’t it B> one of the most
er: /3/ it’s the most A> yeah okay B> but I’m interesting
in drug quality control A> hm: okay well I’ll ask you
about that in a in a few minutes B> ah A> but you you
chose to do the medicine the medical version this morning
B> (laughter) yes A> so I’m going to ask you to have
another look at this could you turn to page one it’s the
very first page in there B> first page yeah A> yeah you
remember reading this this morning very [quickly B> yes]
A> I guess B> very very quickly [(laughter)] A> yeah B>
that was the trick A> okay well this was about plutonium
in the body: and I’m going to have to ask you about this
diagram here er again: in the bottom left hand side left hand corner down here there’s a box with an X in it: my first question is can you identify what X [is B> bones] A> the bones B> yeah A> okay can you tell me how the bones how how it how it’s related to the rest of the diagram I mean what happens to plutonium and how it gets into the bones and so on B> ah: er /2/ plutonium enters the body through inhalation: okay A> hm hm B> I start from the beginning [(laughter A> okay] B> erm /2/ then it goes to the digestive system A> hm hm B> and to the (inaudible) lungs some in the lungs:: /2/ from the lungs it goes er /2/ through the reticuloendothelial system A> hm which is: [what B> erm] /2/ A> does that what does it mean that word B> well ah it’s a system that er /2/ has to do with erm: /5/ fibrocytosis: A> no you’re going to have to explain for me B> erm: consisted for from er: /2/ cells A> hm hm B> that actually are responsible for er: /6/ to take: er foreign material A> hm hm B> out of the body erm /4/ and this is done: er /2/ when they put into the circulation A> okay right B> (laughter) A> I think I’ve got that B> yeah: and er /2/ from that system goes to the blood A> hm hm B> and from the circulation: er /2/ to bones kidneys and the liver A> yes okay how’s erm: you you said that that plutonium was taken into the body through inhalation B> yeah A> yeah: does that mean to say that plutonium is in the air:: [I mean we breathe it in B> there is some] there: yes plutonium exists there in very small quantities very small quantities A> but not

419
enough to damage us: B> oh no I don’t think so A> hm B> except if you live near or next to nuclear reactors where plutonium [(laughter)] A> yes but we don’t have any nuclear reactors in Cyprus B> no A> so this is not a problem that would come up in: would you you would find in Cyprus at all B> no no A> if erm let’s imagine that you did come across a case like this B> yes A> what kind of drug would you treat it with would you treat it with drugs at all: B> drugs A> yes B> I mean do you mean plutonium poisoning A> yeah:: B> the symptoms yes: A> hm B> but not actually not the er: /4/ the real disease A> hm what what would the symptoms be B> oh: it might be symptoms of er: radiation exposure like diarrhoea vomiting A> hm B> headaches and er: /4/ ataxia erm:: /4/ convulsions sweating A> it sounds pretty horrible B> yeah it is [(laughter)] A> okay well I’ll leave that now I I won’t I won’t press you any farther on plutonium poisoning the course you’re going to going to you want to take in England I I presume B> yeah A> yeah what what is it B> it’s a course on I’m not sure because er /7/ it says United Kingdom or or Denmark Denmark /7/ I think A> sorry I I didn’t B> er it was for United Kingdom or Denmark A> or Denmark B> yeah A> so you don’t know which one you’re going to B> no A> do you speak Danish B> no (laughter) A> probably the United Kingdom well unless of course they [teach B> many] teach A> yeah [they might B> (inaudible)] A> teach [it B> yeah] A> in En in En in English in Denmark but but you don’t know where you’re going to do
it B> no not [yet A> no] hm how about the the course
itself quality control: drugs B> yeah A> well what does
it involve B> erm: ah /2/ first we've to control drugs A>
hm B> ah you want me to say what we mean by quality
control [something A> yeah] B> like that A> please B> we
mean ah: we mean /2/ (inaudible) and testers tests which
are used to to determine the density purity potency
quantity and of course quality of pharmaceutical
preparations A> okay I I thought that most pharmaceutical
preparations they were produced by companies B> yeah A>
and then they were sold I mean does this mean to say
that: say you were going to buy a drug from an English
company or a company from the States: but when it came to
Cyprus you you would have to check it first: before you
actually used it in Cyprus [or are B> or yeah] A> you
checking the quality of [drugs produced B> the first
time] A> in Cyprus B> yes A> [I mean erm: B> (inaudible)]
A> quality of what I mean [to say (inaudible) B> I mean
all] pharmaceutical companies have their own: er /2/
quality control departments A> yeah B> they check their
products for their quality but er all products registered
and sold in the Cyprus market: /2/ they are first checked
from their quality and then registered here A> right B> I
mean: for the first time A> that sounds sensible B> yeah
A> hm B> and of course: /4/ I mean the government
(inaudible): er er /2/ has two pharma two: erm /5/ labs
two laboratories the pharmaceutical laboratories and the
chemical state laboratory A> hm B> in the pharmaceutical
laboratory are checked for products for medicines bought by the government for the use of to be used in the hospitals it sounds like very interesting work it is quite fascinating yeah has to do with chemical reactions yeah hm: yes it’s fascinating I think it’s much more fascinating than being involved in this side of hospital work than perhaps you know the medical practitioners going around and seeing patients and things I mean I would find this more fascinating more fascinating with chemical reactions yeah: but are yeah it’s fascinating has to do with photometric methods [and hm:] (inaudible) yeah: okay well I don’t think I need to ask you anything else unless you have anything to ask me (laughter) no (laughter) [okay do you] live in Cyprus yes I do really yes oh I thought you were coming from England well yes I am I am English because all examiners for this have to be English but I’ve lived in Cyprus for: almost nine years now you like it of course I’m more Cypriot than English now (laughter) don’t speak Greek do you yes I do really yes my wife is a Cypriot oh as well and because it’s a very very difficult language not particularly no it’s very rule governed: erm so once you get the hang of the rules you’re okay (laughter) but er: er no I’ve never
I've [never B> (inaudible)] A> yeah I mean it's very it's a nice language but er no I've never worked in England and I feel er I all my working life has been in Cyprus B> oh A> I I know Cyprus better than I know England [(laughter)] anyway er B> did I pass the exams A> can't tell you that [(laughter) B> yeah] A> because of course I don't know erm B> well I mean your part A> oh: (laughter) you're you're a good speaker: I I mean I shouldn't worry too much: but for the rest of it I don't know B> yeah A> I don't mark the rest of it B> thank you very much A> okay bye B> bye

<table>
<thead>
<tr>
<th>Category</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tally</td>
<td>0</td>
<td>18</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Appendix to Chapter Five

Table 5.1. Correlation of error scores

<table>
<thead>
<tr>
<th></th>
<th>Cat1</th>
<th>Cat2</th>
<th>Cat3</th>
<th>Cat4</th>
<th>Cat5</th>
<th>Cat6</th>
<th>Cat7</th>
<th>Cat8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat2</td>
<td>.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cat3</td>
<td>.44</td>
<td>.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cat4</td>
<td>.37</td>
<td>.03</td>
<td>.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cat5</td>
<td>.21</td>
<td>-.26</td>
<td>-.01</td>
<td>.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cat6</td>
<td>.30</td>
<td>.18</td>
<td>.36</td>
<td>.49</td>
<td>.16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cat7</td>
<td>.15</td>
<td>.19</td>
<td>.75</td>
<td>.44</td>
<td>.20</td>
<td>.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cat8</td>
<td>.59</td>
<td>.34</td>
<td>.68</td>
<td>.41</td>
<td>.12</td>
<td>.23</td>
<td>.52</td>
<td></td>
</tr>
<tr>
<td>Cat9</td>
<td>.31</td>
<td>-.11</td>
<td>.46</td>
<td>.25</td>
<td>.24</td>
<td>.44</td>
<td>.36</td>
<td>.41</td>
</tr>
</tbody>
</table>

Table 5.2. Principal Components solution with Varimax Rotation for the initial study into possible category reduction.

Unrotated extraction:

<table>
<thead>
<tr>
<th></th>
<th>Eigenvalue</th>
<th>% of variance explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 1</td>
<td>3.71</td>
<td>41.23</td>
</tr>
<tr>
<td>Factor 2</td>
<td>1.53</td>
<td>16.96</td>
</tr>
</tbody>
</table>

Table 5.3. Multivariate test results in the discriminant analysis of lexical and grammatical accuracy.

<table>
<thead>
<tr>
<th>Categorial variable</th>
<th>Wilks' Lambda</th>
<th>F</th>
<th>df</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band scores</td>
<td>.304</td>
<td>4.34</td>
<td>6,32</td>
<td>.003</td>
</tr>
</tbody>
</table>

1st Discriminant Function $X^2 = 34.59$ $p = .00$

2nd Discriminant Function $X^2 = 9.45$ $p = .14$
Table 5.4. Types of utterances in turns as a percentage of the total number of turns.

<table>
<thead>
<tr>
<th>ELTS Band Score</th>
<th>Total No. of turns</th>
<th>Single Word Turns as % of turns</th>
<th>Fillers as % of single word turns</th>
<th>Backchannels of single word turns</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>91</td>
<td>38.50</td>
<td>2.86</td>
<td>28.57</td>
</tr>
<tr>
<td>6</td>
<td>29</td>
<td>38.00</td>
<td>0.00</td>
<td>9.09</td>
</tr>
<tr>
<td>8</td>
<td>38</td>
<td>18.43</td>
<td>0.00</td>
<td>42.86</td>
</tr>
<tr>
<td>6</td>
<td>58</td>
<td>37.93</td>
<td>0.00</td>
<td>13.64</td>
</tr>
<tr>
<td>6</td>
<td>70</td>
<td>35.71</td>
<td>12.00</td>
<td>28.00</td>
</tr>
<tr>
<td>5</td>
<td>59</td>
<td>38.98</td>
<td>13.04</td>
<td>8.70</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>37.50</td>
<td>6.67</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>63</td>
<td>47.86</td>
<td>3.70</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>52</td>
<td>57.69</td>
<td>46.67</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>32.00</td>
<td>6.25</td>
<td>25.00</td>
</tr>
<tr>
<td>4</td>
<td>45</td>
<td>44.44</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>44</td>
<td>50.00</td>
<td>0.00</td>
<td>13.64</td>
</tr>
<tr>
<td>6</td>
<td>74</td>
<td>45.95</td>
<td>5.88</td>
<td>20.59</td>
</tr>
<tr>
<td>6</td>
<td>95</td>
<td>35.79</td>
<td>0.00</td>
<td>26.47</td>
</tr>
<tr>
<td>6</td>
<td>55</td>
<td>40.00</td>
<td>0.00</td>
<td>50.00</td>
</tr>
<tr>
<td>7</td>
<td>47</td>
<td>8.51</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>39</td>
<td>23.08</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>55</td>
<td>54.55</td>
<td>6.67</td>
<td>20.00</td>
</tr>
<tr>
<td>9</td>
<td>43</td>
<td>20.93</td>
<td>0.00</td>
<td>22.22</td>
</tr>
<tr>
<td>4</td>
<td>33</td>
<td>36.36</td>
<td>25.00</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>53</td>
<td>15.09</td>
<td>12.50</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 5.5. Multivariate test results in the discriminant analysis of single word turns and backchanneling.

<table>
<thead>
<tr>
<th>Categorical variable</th>
<th>Wilks' Lambda</th>
<th>F</th>
<th>df</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band scores</td>
<td>.384</td>
<td>5.22</td>
<td>4,34</td>
<td>.002</td>
</tr>
</tbody>
</table>

Table 5.6. Multivariate test results in the discriminant analysis of use of lexis.

<table>
<thead>
<tr>
<th>Categorical variable</th>
<th>Wilks' Lambda</th>
<th>F</th>
<th>df</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band scores</td>
<td>.107</td>
<td>3.0</td>
<td>15,36</td>
<td>.003</td>
</tr>
</tbody>
</table>
Table 5.7. Mean of total number of words, unique words and word forms by band.

<table>
<thead>
<tr>
<th></th>
<th>Band 1</th>
<th>Band 2</th>
<th>Band 3</th>
<th>Band 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of words</td>
<td>394.57</td>
<td>616.25</td>
<td>632.75</td>
<td>607.50</td>
</tr>
<tr>
<td>Number of unique words</td>
<td>91.00</td>
<td>116.63</td>
<td>140.50</td>
<td>112.50</td>
</tr>
<tr>
<td>Number of word forms</td>
<td>152.29</td>
<td>214.75</td>
<td>238.25</td>
<td>214.50</td>
</tr>
</tbody>
</table>

Table 5.8. Spearman Rho for individual categories of Accuracy and ELTS scores.

<table>
<thead>
<tr>
<th>Category</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELTS Scores</td>
<td>.02</td>
<td>.09</td>
<td>-.26</td>
<td>-.28</td>
<td>-.30</td>
<td>-.03</td>
<td>.11</td>
<td>-.38</td>
<td>-.31</td>
</tr>
</tbody>
</table>

Table 5.9. Means for students scoring different bands for category 1 in the definition of accuracy.

<table>
<thead>
<tr>
<th>Bands</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.86</td>
<td>1.00</td>
<td>0.75</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Table 5.10. Means for students scoring different bands for category 2 in the definition of accuracy.

<table>
<thead>
<tr>
<th>Bands</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.14</td>
<td>2.63</td>
<td>0.75</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 5.11. Means for students scoring different bands for category 3 in the definition of accuracy.

<table>
<thead>
<tr>
<th>Bands</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>4.86</td>
<td>4.75</td>
<td>2.75</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 5.12. Means for students scoring different bands for category 4 in the definition of accuracy.

<table>
<thead>
<tr>
<th>Bands</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.86</td>
<td>5.62</td>
<td>1.50</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 5.13. Means for students scoring different bands for category 5 in the definition of accuracy.

<table>
<thead>
<tr>
<th>Bands</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.57</td>
<td>1.75</td>
<td>0.75</td>
<td>0.00</td>
</tr>
</tbody>
</table>
### Table 5.14. Means for students scoring different bands for category 6 in the definition of accuracy.

<table>
<thead>
<tr>
<th>Bands</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.71</td>
</tr>
<tr>
<td>2</td>
<td>3.75</td>
</tr>
<tr>
<td>3</td>
<td>3.00</td>
</tr>
<tr>
<td>4</td>
<td>1.50</td>
</tr>
</tbody>
</table>

### Table 5.15. Means for students scoring different bands for category 7 in the definition of accuracy.

<table>
<thead>
<tr>
<th>Bands</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.57</td>
</tr>
<tr>
<td>2</td>
<td>2.25</td>
</tr>
<tr>
<td>3</td>
<td>2.00</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
</tr>
</tbody>
</table>

### Table 5.16. Means for students scoring different bands for category 8 in the definition of accuracy.

<table>
<thead>
<tr>
<th>Bands</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.29</td>
</tr>
<tr>
<td>2</td>
<td>3.13</td>
</tr>
<tr>
<td>3</td>
<td>2.00</td>
</tr>
<tr>
<td>4</td>
<td>1.00</td>
</tr>
</tbody>
</table>

### Table 5.17. Means for students scoring different bands for category 9 in the definition of accuracy.

<table>
<thead>
<tr>
<th>Bands</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.57</td>
</tr>
<tr>
<td>2</td>
<td>1.38</td>
</tr>
<tr>
<td>3</td>
<td>0.50</td>
</tr>
<tr>
<td>4</td>
<td>0.50</td>
</tr>
</tbody>
</table>
Table 5.18. Rank orders on 7 studies for five grammatical morphemes.

<table>
<thead>
<tr>
<th>Study</th>
<th>Plural (-s)</th>
<th>Progressive (-ing)</th>
<th>Past irr.</th>
<th>Possessive</th>
<th>3rd person (-s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>de Villiers &amp; de Villiers (1973)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Dulay &amp; Burt (1973)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Dulay &amp; Burt (1974)</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Bailey et. al (1974)</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Larsen-Freeman (1975a)</td>
<td>1</td>
<td></td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Larsen-Freeman (1975b)</td>
<td>1</td>
<td></td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Fulcher</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 5.19. Disconfirmation matrix for eight aspects of grammar.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Pronoun case</td>
<td>-</td>
<td>0</td>
<td>9.52</td>
<td>4.76</td>
<td>4.76</td>
<td>7.14</td>
<td>0</td>
<td>8.33</td>
</tr>
<tr>
<td>2. contracted copula</td>
<td>19.05</td>
<td>0</td>
<td>-</td>
<td>4.76</td>
<td>0</td>
<td>6.66</td>
<td>0</td>
<td>18.18</td>
</tr>
<tr>
<td>3. plural (-es)</td>
<td>25.00</td>
<td>0</td>
<td>18.18</td>
<td>7.69</td>
<td>8.33</td>
<td>0</td>
<td>11.11</td>
<td>-</td>
</tr>
<tr>
<td>4. past irr.</td>
<td>33.33</td>
<td>0</td>
<td>20.00</td>
<td>13.33</td>
<td>6.66</td>
<td>-</td>
<td>9.09</td>
<td>25.00</td>
</tr>
<tr>
<td>5. plural (-s)</td>
<td>28.57</td>
<td>4.76</td>
<td>23.81</td>
<td>-</td>
<td>9.52</td>
<td>13.33</td>
<td>6.25</td>
<td>15.38</td>
</tr>
<tr>
<td>6. article</td>
<td>61.91</td>
<td>11.76</td>
<td>47.62</td>
<td>42.86</td>
<td>-</td>
<td>33.33</td>
<td>7.14</td>
<td>66.66</td>
</tr>
<tr>
<td>7. progressive (-ing)</td>
<td>52.94</td>
<td>-</td>
<td>47.06</td>
<td>32.29</td>
<td>11.76</td>
<td>30.77</td>
<td>16.66</td>
<td>60.00</td>
</tr>
<tr>
<td>8. 3rd person (-s)</td>
<td>60.00</td>
<td>16.67</td>
<td>40.00</td>
<td>43.75</td>
<td>14.29</td>
<td>9.09</td>
<td>-</td>
<td>85.96</td>
</tr>
</tbody>
</table>

429
Table 5.20. Stair matrix for eight aspects of grammar

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. pronoun case</td>
<td>(10)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>2. contracted copula</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. plural (-es)</td>
<td>+</td>
<td>(10)</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. past irr.</td>
<td></td>
<td>+</td>
<td>(10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. plural (-s)</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. article</td>
<td></td>
<td></td>
<td></td>
<td>(10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. progressive (-ing)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. 3rd person (-s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.21. A sample coded transcript.

In the following transcript, the categories of error listed on pages 191 - 197 are given between slashes. The notation "/3/" would therefore indicate an error in tense. "A>" indicates the speech of the interviewer, "B>" the speech of the student and "C>" the speech of a second rater.

Utterances enclosed by square brackets indicates speech overlap, while round brackets indicate noises, the removal of identifying information or sections of inaudible speech.

The oral interview in which this sample was collected consisted of an introductory phase in which the student was put at his/her ease through the use of general questions about home or work, followed by
questions on a text which had been read previously. After this, students are asked about their future plans, followed by a brief wind-up, mainly involving leave taking.

**Interview 5: ELTS General Academic Module: Band 4**

A> I'm (NAME) I'll introduce you: (NAME) C> how do you do A> would you like to have a seat: well: how did you find the test this morning B> er it's /3/ : er: difficult: A> difficult: B> yes: A> how long have you been studying English B> er I study /3/ for er: ten or twelve years A> ten or twelve years at is that at: at school or B> er no outside the /8/ school A> outside of school you took private lessons:: (inaudible) B> /6/ /8/ private institute A> hm hm B> for ten or twelve years: A> you're not still at school you look older B> no I am in the army now B> you are in the army where abouts are you: B> I was in Limassol at Polemedia A> Polemedia: you were are you B> no I don't A> where are you now: B> now I am in the army: A> yes but whereabouts are you based: where do you work in the army now B> I work in the /8/ Polemedia er: PYROVOLIKO A> the artillery B> yes A> yeah: is this a job you chose:: were you just put there: B> I don't understand A> did you choose: to go into the artillery B> yes yes: I choose /3/ to go there because /5/ is near to my house: and er there are more: friends: there A> yeah: what kind of things do you have to do every day B> er: er I work I work in the office in the er in the army: so
the: it's the same thing every day the same routine: er I’ve got to go every morning to the office: er: I do the work er: that my officer: showed /3/ to me to do::: the same thing every day A> yeah: do you get lots of free time B> yes in the afternoon I didn’t /3/ have anything to do: I play basketball: in the army A> hm hm B> er football: or do doing /3/ some er: exercise /7/ in the fields A> hm hm: oh it it sounds as if life in the army is getting better: do you enjoy being in the army or B> I enjoy to be /9/ in the army but er: you don’t er: have the /8/ many /9/ : er time to go out to your home A> hm: B> so that is to that isn’t very good A> when when do you finish: when will you leave the army B> yes: A> when will you [leave the army B> er when the] er about a year A> you mean there’s another year to go B> yes:: A> right you remember reading this booklet earlier this morning: may I ask you to have another look at it::: can you turn to page six::: do you remember reading this passage this morning and you had to answer some questions on it: yes: right I’m not going to ask you to read any of it again just to it’s just to remind you (inaudible) the page: can you remember anything about what the passage said about crime rate: the crime rate amongst girls B> er: I think er: it it’s er talking /4/ about the girls who is /3/ er: who are putting /3/ them /2/ behind the /8/ bars:: and er:: they do some er house er: housework: house in the /8/ Bullwood Hall::: A> is the crime rate amongst girls: presumably this is in England is it increasing or
decreasing do you think: B> I think er here it er increased: /3/ A> quickly B> er not quickly A> not quickly it's going up fairly slowly B> yes A> yeah: can you think of any reasons why the crime rate amongst girls is increasing at this speed: B> it is increasing because er: the family of the: the families of the girls had er: are not er show /3/ them the right er way of life: and er: they don't have any works /7/ to do: so the this /2/ people er: take the get /4/ to the this bad way of life: so the the crime inc increases A> hm so you think it's a: it starts with the family B> yes A> in this case yeah: how about in Cyprus: do you think there's a problem of crime amongst girls in Cyprus: B> there isn't a big er crime (inaudible) I think there is a small A> it's small B> some er village /7/ in er::: A> what if there's a small amount of crime: amongst girls: what kind of crime do you think it is:: B> the the:: er the they get to do very er: heavy works /7/ : er in the fields er: the house:: they they have er they have their work they gets /3/ like er: like er: an animal /7/ A> hm B> some villages not er: not er: town centres:: A> if er if you were in the police force for example: what kind of action would you take to reduce the rate of crime: B> I will /9/ er arrested /3/ the the that /2/ people who do /4/ that /2/ er crime A> hm hm B> and I will /9/ er: to not /9/ er: no more er crime: not increases /3/ the the crime in the /8/ Cyprus A> hm: but it's not a very large problem anyway [in Cyprus B> no no] A> it's fairly small I would
have thought:: anyway let’s finish with the books okay er (inaudible) (COUGH) er let me ask you something: why are you taking this test: what what are you thinking of doing in the future B> /5/ take /3/ the test because I want the: want to go to the /8/ university in England A> hm hm B> so I want to to show to the university that I have done: passed the English er test A> er which university are you thinking of going to: B> I think er I’ll go to the /8/ Kent University A> Kent B> yes A> and what are you going to study: B> er computer and business studies computer and administration together A> why did you choose that B> because I think that er it’s the most er: it’s the best work to do in Cyprus to er: to er: to have a good to find a /8/ work in Cyprus A> so so there there are lots of positions available B> yes A> do you want to work for a company or when you come back to Cyprus B> no er my father have /3/ er: some er company /7/ : and I will work with er: I will continue the job of my father A> so everything ready when you come back: okay right I think we can stop there: thank you very much for coming it’s very nice to meet you B> good bye A> good bye C> bye bye

<table>
<thead>
<tr>
<th>Category</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error tally</td>
<td>0</td>
<td>4</td>
<td>15</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

434
Appendix to Chapter Six

Material used in the Three Oral Testing Tasks
Task
One
Task
One
Population and Poverty

We can define "land" as the natural resources which provide man with food and raw materials, and we cannot escape the fact that natural resources are limited in relation to rapidly growing populations.

In some countries the standard of living is the same as it was centuries ago. These countries have increased their output, but their populations have increased by the same amount, so that the gross national product (GNP) per capita (or per person) has declined. If populations increase by the same amount as total output, there is no increase in per capita income. If populations increase faster than output, per capita income falls.

For example, suppose Holland has a birth rate of sixteen per thousand, and in Ethiopia the figure is fifty-one. If the death rate in Holland is ten, and in Ethiopia twenty-five, the net increase in Holland is six per thousand (0.6%). In Ethiopia it is twenty-six per thousand (2.6%). If both countries' economies grow by 2% per annum, Holland's standard of living will increase per capita, whereas Ethiopia's per capita standard of living will be falling.

This is one of the factors which make the gap between developed and developing countries so wide, and there is a danger that this gap will continue to widen.

One solution would be to increase the productivity of land. It is often difficult to do this. There may be climatic reasons for the difficulty, or there may not be enough capital employed to produce more. But whatever happened, economic growth would have to be greater than population growth if per capita income were to increase. Although countries in the Third World may take other steps which would help alleviate the problems which poverty creates, this is why controlling population growth is so important. A proper population policy is necessary if economic development is to take place.
Fig. 1: Population

Fig. 2: Birth and Death Rates

442
Spare him more than just a thought.

Just ten pounds a month can help a child and its community provide themselves with life's basics like fresh water, a healthy diet and basic education.

Sponsor a child through Britain's biggest sponsorship charity and you'll both see big results — solid mud dams, good harvests, strong school buildings. A photo and regular reports will keep you in touch. If you wish you can even write and receive letters. It's almost like they're one of the family.

Every new sponsor gives another child the chance of a life worth living.

**SPONSOR A CHILD, PLEASE**

- □ Yes, I would like to sponsor a child and enclose my first contribution.
- □ £10 (monthly) □ £120 (yearly)
- □ I can't sponsor a child, but enclose a gift of
  - □ £200 □ £100 □ £50 □ £25 □ £

- □ Please send more details.

**Name Mr/Ms**

**Address**

**Postcode**

**Tel**

Make cheques/P.O.s payable to ActionAid. Send to: The Rt. Hon. Christopher Chataway, ActionAid, Tapstone Road, Chard, Somerset TA20 2AB.
INTRODUCTION

Cyprus has a well organised and effective system of primary, secondary and partly of further education but until recently it did not have any accredited institutions of higher education that could meet even the most vital needs of the country in this domain.

These needs were met through the attendance of students from Cyprus at various Universities abroad. The desire of the people of Cyprus for higher education can be seen from the fact that though Cyprus lacked its own universities it was, according to recent official statistics, one of the countries with, proportionately, the highest number of University graduates worldwide.

Among the Greek population of Cyprus there were in the last few years about ten thousand students studying abroad. In the academic year 1988-89 there were 9,410 such students. Their distribution in various countries was as follows: Greece 40 per cent; United Kingdom 22 per cent; United States 20 per cent; Federal Republic of Germany 7 per cent; Italy 3 per cent; France 2 per cent; Austria 2 per cent; Eastern European countries 3 per cent and other countries 1 per cent. Unfortunately we have no corresponding data for Turkish Cypriot students studying abroad.

It is obvious that Cyprus can easily sustain its own University and that a Cyprus University will have beneficial consequences for the country not only in the economic but also in the cultural and political fields.

In light of the above considerations the Government of the Republic of Cyprus has established by law 144 of 1989 the University of Cyprus. The law was drafted on the basis of a Report by a Preparatory Committee in which eighteen University Professors and other specialists from Cyprus took part. The Committee proposed the establishment of an academically autonomous University of high calibre in order to promote knowledge and learning and thus contribute to the social, economic and cultural development of the people of Cyprus as a whole. It is also expected that it will contribute towards the mutual understanding of the communities of Cyprus, in the cultivation of their several cultures and traditions and in the consolidation of the unity of the Cyprus state.

The Council of Ministers has subsequently appointed a nine-member Interim Governing Board as well as a Secretariat for planning and implementing the work required for the operation of the University within the stipulated timetable. The Interim Governing Board is performing the duties of the Council and the Senate of the University until the time that these bodies are duly constituted.

It is envisaged that the University will admit its first 440 students in September 1992. The annual expected increase in the student body is as follows: 440 (1992-3), 900 (1993-4), 1,400 (1994-5), 1,900 (1995-6). In its subsequent steady state the University will have a total of 4,000 students.
Task Three

Task Card: Group Discussion.

You will discuss the subject of "Education in Cyprus" for ten to fifteen minutes. You have ten minutes to prepare.

Here are some ideas for topics which you may discuss if you wish. You may talk about other topics:

* Quality of School education/private education.
* Importance of education for Cypriot students.
* Is there a need for a University in Cyprus? Why?
* Where should the University be built?
* Should the language of tuition be Greek and Turkish, or English? What are the advantages/disadvantages?
* Who will attend the University?
* How does education affect the Cypriot economy?
Appendix to Chapter Seven

Table 7.1. Varimax Rotated Factor Analysis of the Questionnaire for Task 1.

<table>
<thead>
<tr>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>.23</td>
<td>-.25</td>
<td>-.06</td>
</tr>
<tr>
<td>Q2</td>
<td>.07</td>
<td>.70</td>
<td>.21</td>
</tr>
<tr>
<td>Q3</td>
<td>.07</td>
<td>.71</td>
<td>.08</td>
</tr>
<tr>
<td>Q4</td>
<td>.02</td>
<td>.06</td>
<td>.05</td>
</tr>
<tr>
<td>Q5</td>
<td>.05</td>
<td>.60</td>
<td>.22</td>
</tr>
<tr>
<td>Q6</td>
<td>1.00</td>
<td>-.02</td>
<td>.13</td>
</tr>
<tr>
<td>Q7</td>
<td>.07</td>
<td>-.39</td>
<td>.29</td>
</tr>
<tr>
<td>Q8</td>
<td>-.07</td>
<td>.02</td>
<td>.82</td>
</tr>
<tr>
<td>Q9</td>
<td>.04</td>
<td>-.32</td>
<td>-.01</td>
</tr>
<tr>
<td>Q10</td>
<td>.08</td>
<td>.41</td>
<td>-.32</td>
</tr>
<tr>
<td>Q11</td>
<td>-.07</td>
<td>.26</td>
<td>-.21</td>
</tr>
<tr>
<td>Q12</td>
<td>.03</td>
<td>.17</td>
<td>.03</td>
</tr>
<tr>
<td>Q13</td>
<td>.05</td>
<td>.13</td>
<td>.65</td>
</tr>
<tr>
<td>Q14</td>
<td>-.07</td>
<td>-.04</td>
<td>-.12</td>
</tr>
<tr>
<td>Q15</td>
<td>.24</td>
<td>.16</td>
<td>.23</td>
</tr>
</tbody>
</table>

Variance explained by rotated factors:
1.99 1.93 1.51 1.21

% of total variance explained:
13.28 12.84 10.07 8.09 44.28

Table 7.2. Varimax Rotated Factor Analysis of the Questionnaire for Task 2.

<table>
<thead>
<tr>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>.00</td>
<td>.76</td>
<td>.17</td>
</tr>
<tr>
<td>Q2</td>
<td>-.24</td>
<td>.03</td>
<td>.83</td>
</tr>
<tr>
<td>Q3</td>
<td>.06</td>
<td>.17</td>
<td>.61</td>
</tr>
<tr>
<td>Q4</td>
<td>.35</td>
<td>-.06</td>
<td>.04</td>
</tr>
<tr>
<td>Q5</td>
<td>-.14</td>
<td>-.07</td>
<td>.16</td>
</tr>
<tr>
<td>Q6</td>
<td>.55</td>
<td>.41</td>
<td>-.01</td>
</tr>
<tr>
<td>Q7</td>
<td>-.01</td>
<td>.20</td>
<td>-.15</td>
</tr>
<tr>
<td>Q8</td>
<td>.80</td>
<td>.22</td>
<td>-.20</td>
</tr>
<tr>
<td>Q9</td>
<td>.69</td>
<td>-.22</td>
<td>-.13</td>
</tr>
<tr>
<td>Q10</td>
<td>.08</td>
<td>-.06</td>
<td>.11</td>
</tr>
<tr>
<td>Q11</td>
<td>-.27</td>
<td>.25</td>
<td>.22</td>
</tr>
<tr>
<td>Q12</td>
<td>-.59</td>
<td>.20</td>
<td>.05</td>
</tr>
<tr>
<td>Q13</td>
<td>.44</td>
<td>.29</td>
<td>.13</td>
</tr>
<tr>
<td>Q14</td>
<td>-.75</td>
<td>-.05</td>
<td>.40</td>
</tr>
<tr>
<td>Q15</td>
<td>.71</td>
<td>.04</td>
<td>.17</td>
</tr>
</tbody>
</table>

Variance explained by rotated factors:
3.30 1.12 1.44 1.03

% of total variance explained:
21.97 7.45 9.62 6.85 45.89

446
Table 7.3. Varimax Rotated Factor Analysis of the Questionnaire for Task 3.

<table>
<thead>
<tr>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>.33</td>
<td>-.02</td>
<td>-.13</td>
</tr>
<tr>
<td>Q2</td>
<td>.02</td>
<td>.85</td>
<td>.04</td>
</tr>
<tr>
<td>Q3</td>
<td>-.08</td>
<td>.74</td>
<td>.04</td>
</tr>
<tr>
<td>Q4</td>
<td>.54</td>
<td>-.45</td>
<td>.22</td>
</tr>
<tr>
<td>Q5</td>
<td>-.02</td>
<td>.69</td>
<td>-.04</td>
</tr>
<tr>
<td>Q6</td>
<td>.52</td>
<td>-.10</td>
<td>-.03</td>
</tr>
<tr>
<td>Q7</td>
<td>.17</td>
<td>-.19</td>
<td>.56</td>
</tr>
<tr>
<td>Q8</td>
<td>.91</td>
<td>-.12</td>
<td>.02</td>
</tr>
<tr>
<td>Q9</td>
<td>.15</td>
<td>-.23</td>
<td>.04</td>
</tr>
<tr>
<td>Q10</td>
<td>.35</td>
<td>.09</td>
<td>.22</td>
</tr>
<tr>
<td>Q11</td>
<td>-.13</td>
<td>.16</td>
<td>.73</td>
</tr>
<tr>
<td>Q12</td>
<td>-.55</td>
<td>.35</td>
<td>.10</td>
</tr>
<tr>
<td>Q13</td>
<td>.78</td>
<td>.09</td>
<td>-.07</td>
</tr>
<tr>
<td>Q14</td>
<td>-.57</td>
<td>.34</td>
<td>.19</td>
</tr>
<tr>
<td>Q15</td>
<td>.70</td>
<td>.05</td>
<td>.16</td>
</tr>
</tbody>
</table>

Variance explained by rotated factors:

<table>
<thead>
<tr>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.41</td>
<td>2.35</td>
<td>1.04</td>
<td>1.46</td>
</tr>
</tbody>
</table>

% of total variance explained:

<table>
<thead>
<tr>
<th>Task</th>
<th>Question Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2.82</td>
<td>1.03</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>5</td>
<td>3.51</td>
<td>1.29</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>5</td>
<td>2.87</td>
<td>1.31</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>5</td>
<td>3.60</td>
<td>0.65</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>4</td>
<td>2.36</td>
<td>1.03</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>5</td>
<td>3.20</td>
<td>1.08</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>4</td>
<td>2.51</td>
<td>0.79</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>5</td>
<td>3.91</td>
<td>0.79</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>5</td>
<td>4.42</td>
<td>0.66</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>4</td>
<td>3.20</td>
<td>0.99</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>5</td>
<td>2.09</td>
<td>0.95</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>4</td>
<td>1.71</td>
<td>0.70</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>5</td>
<td>3.89</td>
<td>0.83</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>4</td>
<td>1.69</td>
<td>0.63</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>5</td>
<td>3.98</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 7.4. Summary Statistics for the Questionnaires

<table>
<thead>
<tr>
<th>Task</th>
<th>Question Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>5</td>
<td>3.13</td>
<td>1.06</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>5</td>
<td>2.53</td>
<td>1.12</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>5</td>
<td>2.58</td>
<td>1.28</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>3.09</td>
<td>0.76</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>5</td>
<td>2.49</td>
<td>1.04</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>4</td>
<td>2.69</td>
<td>0.95</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>5</td>
<td>2.98</td>
<td>1.18</td>
</tr>
</tbody>
</table>
Table 7.4 continued

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>4</th>
<th>3.89</th>
<th>0.94</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>5</td>
<td>2.93</td>
<td>1.03</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>3</td>
<td>2.07</td>
<td>0.86</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>5</td>
<td>3.09</td>
<td>1.04</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>5</td>
<td>3.60</td>
<td>0.92</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>5</td>
<td>2.31</td>
<td>0.97</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>5</td>
<td>2.78</td>
<td>1.02</td>
</tr>
</tbody>
</table>

---

Table 7.5. ANOVA and Tukey HSD results for task 1, question 5

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Df</th>
<th>Mean Square</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>7.53</td>
<td>2</td>
<td>3.76</td>
<td>4.08</td>
<td>.02</td>
</tr>
<tr>
<td>Within Groups</td>
<td>38.78</td>
<td>42</td>
<td>0.92</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Matrix of Mean Differences

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Matrix of Probabilities

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No. of Cases</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

448
Table 7.6. ANOVA and Tukey HSD results for task 1, question 6

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Df</th>
<th>Mean Square</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>15.66</td>
<td>2</td>
<td>7.83</td>
<td>9.25</td>
<td>.00</td>
</tr>
<tr>
<td>Within Groups</td>
<td>35.54</td>
<td>42</td>
<td>0.85</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Matrix of Mean Differences

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.76</td>
<td>0.11</td>
<td>0.78</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Matrix of Probabilities

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.11</td>
<td>0.04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No. of Cases</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>10</td>
<td>2.30</td>
</tr>
<tr>
<td>Group 2</td>
<td>17</td>
<td>3.06</td>
</tr>
<tr>
<td>Group 3</td>
<td>18</td>
<td>3.83</td>
</tr>
</tbody>
</table>

Table 7.7. ANOVA and Tukey HSD results for task 2, question 3

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Df</th>
<th>Mean Square</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>18.50</td>
<td>2</td>
<td>9.25</td>
<td>7.13</td>
<td>.00</td>
</tr>
<tr>
<td>Within Groups</td>
<td>54.48</td>
<td>42</td>
<td>1.30</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Matrix of Mean Differences

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.78</td>
<td>0.21</td>
<td>0.78</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Matrix of Probabilities

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.21</td>
<td>0.30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No. of Cases</th>
<th>Mean</th>
<th>Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>10</td>
<td>2.60</td>
</tr>
<tr>
<td>Group 2</td>
<td>17</td>
<td>1.82</td>
</tr>
<tr>
<td>Group 3</td>
<td>18</td>
<td>3.28</td>
</tr>
</tbody>
</table>

449
Table 7.8. ANOVA and Tukey HSD results for task 2, question 13

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Df</th>
<th>Mean Square</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>6.59</td>
<td>2</td>
<td>3.30</td>
<td>4.58</td>
<td>.02</td>
</tr>
<tr>
<td>Within Groups</td>
<td>30.21</td>
<td>42</td>
<td>0.72</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Matrix of Mean Differences

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 2</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 2</td>
<td>0.81</td>
<td>0.18</td>
<td>0.06</td>
<td>0.80</td>
</tr>
<tr>
<td>Group 3</td>
<td>0.99</td>
<td>0.18</td>
<td>0.01</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Matrix of Probabilities

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>0.06</td>
<td>0.01</td>
<td>0.06</td>
</tr>
<tr>
<td>Group 2</td>
<td>0.01</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>No. of Cases</td>
<td>Mean</td>
<td>Standard Deviation</td>
<td></td>
</tr>
<tr>
<td>Group 1</td>
<td>10</td>
<td>2.90</td>
<td>1.29</td>
</tr>
<tr>
<td>Group 2</td>
<td>17</td>
<td>3.71</td>
<td>0.77</td>
</tr>
<tr>
<td>Group 3</td>
<td>18</td>
<td>3.89</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Table 7.9. Summary statistics for groups on task 2, questions 8 and 15.

<table>
<thead>
<tr>
<th>Question 8</th>
<th>No of cases</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>10</td>
<td>3.00</td>
<td>1.33</td>
</tr>
<tr>
<td>Group 2</td>
<td>17</td>
<td>2.71</td>
<td>1.21</td>
</tr>
<tr>
<td>Group 3</td>
<td>18</td>
<td>3.22</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Question 15

| Group 1    | 10          | 2.30 | 1.06               |
| Group 2    | 17          | 3.41 | 0.51               |
| Group 3    | 18          | 3.39 | 0.61               |

Table 7.10. Task preferences by group

<table>
<thead>
<tr>
<th>No. of Cases</th>
<th>Task 1 %</th>
<th>Task 2 %</th>
<th>Task 3 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>10</td>
<td>4</td>
<td>40</td>
</tr>
<tr>
<td>Group 2</td>
<td>17</td>
<td>5</td>
<td>29.4</td>
</tr>
<tr>
<td>Group 3</td>
<td>18</td>
<td>7</td>
<td>38.9</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
<td>16</td>
<td>35.6</td>
</tr>
</tbody>
</table>
Table 7.11. Manova results for Group (ability level)

### Univariate F Tests

<table>
<thead>
<tr>
<th>Variable</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency, Task 1</td>
<td>440.84</td>
<td>2</td>
<td>220.42</td>
<td>26.60</td>
<td>.00</td>
</tr>
<tr>
<td>Error</td>
<td>298.33</td>
<td>36</td>
<td>8.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluency, Task 2</td>
<td>454.42</td>
<td>2</td>
<td>227.21</td>
<td>27.57</td>
<td>.00</td>
</tr>
<tr>
<td>Error</td>
<td>296.68</td>
<td>36</td>
<td>8.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluency, Task 3</td>
<td>444.23</td>
<td>2</td>
<td>222.16</td>
<td>37.95</td>
<td>.00</td>
</tr>
<tr>
<td>Error</td>
<td>210.73</td>
<td>36</td>
<td>5.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy, Task 1</td>
<td>271.67</td>
<td>2</td>
<td>135.83</td>
<td>22.98</td>
<td>.00</td>
</tr>
<tr>
<td>Error</td>
<td>212.76</td>
<td>36</td>
<td>5.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy, Task 2</td>
<td>288.91</td>
<td>2</td>
<td>144.46</td>
<td>19.26</td>
<td>.00</td>
</tr>
<tr>
<td>Error</td>
<td>269.95</td>
<td>36</td>
<td>7.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy, Task 3</td>
<td>165.72</td>
<td>2</td>
<td>82.86</td>
<td>0.31</td>
<td>.74</td>
</tr>
<tr>
<td>Error</td>
<td>9774.60</td>
<td>36</td>
<td>271.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELTS, Task 1</td>
<td>377.07</td>
<td>2</td>
<td>188.54</td>
<td>25.91</td>
<td>.00</td>
</tr>
<tr>
<td>Error</td>
<td>261.99</td>
<td>36</td>
<td>7.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELTS, Task 2</td>
<td>375.29</td>
<td>2</td>
<td>187.65</td>
<td>20.29</td>
<td>.00</td>
</tr>
<tr>
<td>Error</td>
<td>332.98</td>
<td>36</td>
<td>9.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELTS, Task 3</td>
<td>278.153</td>
<td>2</td>
<td>139.08</td>
<td>18.65</td>
<td>.00</td>
</tr>
<tr>
<td>Error</td>
<td>268.446</td>
<td>36</td>
<td>7.46</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Multivariate Test Statistics

- Wilks’ Lambda = 0.14
  - $F = 5.23$  
  - $Df = 18, 56$  
  - $p = 0.00$
Table 7.12. Manova results for Task preference

<table>
<thead>
<tr>
<th>Variable</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency, Task 1</td>
<td>3.90</td>
<td>2</td>
<td>1.95</td>
<td>0.24</td>
<td>.79</td>
</tr>
<tr>
<td>Error</td>
<td>298.33</td>
<td>36</td>
<td>8.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluency, Task 2</td>
<td>1.96</td>
<td>2</td>
<td>0.98</td>
<td>0.119</td>
<td>.88</td>
</tr>
<tr>
<td>Error</td>
<td>296.68</td>
<td>36</td>
<td>8.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>210.73</td>
<td>36</td>
<td>5.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy, Task 1</td>
<td>5.04</td>
<td>2</td>
<td>2.52</td>
<td>0.43</td>
<td>.66</td>
</tr>
<tr>
<td>Error</td>
<td>212.76</td>
<td>36</td>
<td>5.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy, Task 2</td>
<td>13.79</td>
<td>2</td>
<td>6.89</td>
<td>0.92</td>
<td>.41</td>
</tr>
<tr>
<td>Error</td>
<td>269.95</td>
<td>36</td>
<td>7.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy, Task 3</td>
<td>221.76</td>
<td>2</td>
<td>110.88</td>
<td>0.41</td>
<td>.67</td>
</tr>
<tr>
<td>Error</td>
<td>9774.60</td>
<td>36</td>
<td>271.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELTS, Task 1</td>
<td>1.21</td>
<td>2</td>
<td>0.60</td>
<td>0.08</td>
<td>.92</td>
</tr>
<tr>
<td>Error</td>
<td>261.99</td>
<td>36</td>
<td>7.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELTS, Task 2</td>
<td>6.36</td>
<td>2</td>
<td>3.18</td>
<td>0.34</td>
<td>.71</td>
</tr>
<tr>
<td>Error</td>
<td>332.98</td>
<td>36</td>
<td>9.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELTS, Task 3</td>
<td>0.22</td>
<td>2</td>
<td>0.11</td>
<td>0.01</td>
<td>.97</td>
</tr>
<tr>
<td>Error</td>
<td>268.446</td>
<td>36</td>
<td>7.46</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multivariate Test Statistics

Wilks' Lambda = 0.33

F = 2.32  Df = 18, 56  p = .01
### Table 7.13. Manova results for Group x Task preference

<table>
<thead>
<tr>
<th>Variable</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency, Task 1</td>
<td>30.80</td>
<td>4</td>
<td>7.70</td>
<td>0.93</td>
<td>.46</td>
</tr>
<tr>
<td>Error</td>
<td>298.33</td>
<td>36</td>
<td>8.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluency, Task 2</td>
<td>110.47</td>
<td>4</td>
<td>27.62</td>
<td>3.35</td>
<td>.02</td>
</tr>
<tr>
<td>Error</td>
<td>296.68</td>
<td>36</td>
<td>8.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluency, Task 3</td>
<td>138.19</td>
<td>4</td>
<td>34.55</td>
<td>5.90</td>
<td>.00</td>
</tr>
<tr>
<td>Error</td>
<td>210.73</td>
<td>36</td>
<td>5.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy, Task 1</td>
<td>49.09</td>
<td>4</td>
<td>12.27</td>
<td>2.08</td>
<td>.10</td>
</tr>
<tr>
<td>Error</td>
<td>212.76</td>
<td>36</td>
<td>5.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy, Task 2</td>
<td>104.64</td>
<td>4</td>
<td>26.16</td>
<td>3.49</td>
<td>.02</td>
</tr>
<tr>
<td>Error</td>
<td>269.95</td>
<td>36</td>
<td>7.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy, Task 3</td>
<td>159.73</td>
<td>4</td>
<td>39.93</td>
<td>0.15</td>
<td>.96</td>
</tr>
<tr>
<td>Error</td>
<td>9774.60</td>
<td>36</td>
<td>271.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELTS, Task 1</td>
<td>53.09</td>
<td>4</td>
<td>13.27</td>
<td>1.82</td>
<td>.15</td>
</tr>
<tr>
<td>Error</td>
<td>261.99</td>
<td>36</td>
<td>7.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELTS, Task 2</td>
<td>140.76</td>
<td>4</td>
<td>35.19</td>
<td>3.81</td>
<td>.01</td>
</tr>
<tr>
<td>Error</td>
<td>332.98</td>
<td>36</td>
<td>9.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELTS, Task 3</td>
<td>111.32</td>
<td>4</td>
<td>27.83</td>
<td>3.73</td>
<td>.01</td>
</tr>
<tr>
<td>Error</td>
<td>268.446</td>
<td>36</td>
<td>7.46</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Multivariate Test Statistics

Wilks' Lambda = 0.23

\[ F = 1.42 \quad \text{Df} = 36, 106 \quad p = .09 \]
Table 7.14. Means and Standard Deviations for scores on the Fluency Scale, Accuracy Scale and ELTS scale, by Group and Task preference, for Task 2.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Group</th>
<th>Task Preference</th>
<th>Mean</th>
<th>Sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>1</td>
<td>1</td>
<td>5.00</td>
<td>.94</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>5.13</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>4.56</td>
<td>.66</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>2.75</td>
<td>.79</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>2.75</td>
<td>.71</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>3.93</td>
<td>.87</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>2.68</td>
<td>.53</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>2.88</td>
<td>.18</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td>2.17</td>
<td>.53</td>
</tr>
<tr>
<td>Accuracy</td>
<td>1</td>
<td>1</td>
<td>3.88</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>4.38</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>3.69</td>
<td>.72</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>2.10</td>
<td>.91</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>2.63</td>
<td>.53</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>3.40</td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>2.11</td>
<td>.48</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>2.75</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td>1.97</td>
<td>.48</td>
</tr>
<tr>
<td>ELTS</td>
<td>1</td>
<td>1</td>
<td>6.88</td>
<td>.53</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>6.88</td>
<td>.53</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>6.44</td>
<td>.75</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>4.60</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>4.75</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>5.93</td>
<td>.80</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>4.50</td>
<td>.52</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>5.25</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td>4.17</td>
<td>.77</td>
</tr>
</tbody>
</table>

Table 7.15. Means and Standard Deviations for scores on the Fluency Scale by Group and Task preference, for Task 3.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Group</th>
<th>Task Preference</th>
<th>Mean</th>
<th>Sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>1</td>
<td>1</td>
<td>5.31</td>
<td>.63</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>5.25</td>
<td>.71</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>5.25</td>
<td>.54</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>2.80</td>
<td>.60</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>3.00</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>4.39</td>
<td>.65</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>3.29</td>
<td>.42</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>3.00</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td>2.89</td>
<td>.76</td>
</tr>
</tbody>
</table>
Table 7.16. Correlation coefficients between student assessment of abilities and Rasch ability estimates

<table>
<thead>
<tr>
<th>Rasch estimates of abilities</th>
<th>Fluency</th>
<th>Accuracy</th>
<th>General Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>0.09</td>
<td>0.13</td>
<td>0.01</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.09</td>
<td>0.11</td>
<td>0.02</td>
</tr>
<tr>
<td>General Proficiency</td>
<td>0.08</td>
<td>0.09</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 7.17. Correlation coefficients between reports of anxiety and Rasch estimates of ability.

<table>
<thead>
<tr>
<th>Question 2, Task 1</th>
<th>Fluency</th>
<th>Accuracy</th>
<th>ELTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question 3, Task 1</td>
<td>-0.08</td>
<td>-0.10</td>
<td>-0.07</td>
</tr>
<tr>
<td>Question 5, Task 1</td>
<td>-0.28</td>
<td>-0.25</td>
<td>-0.29</td>
</tr>
<tr>
<td>Question 2, Task 2</td>
<td>-0.10</td>
<td>-0.12</td>
<td>-0.15</td>
</tr>
<tr>
<td>Question 3, Task 2</td>
<td>-0.09</td>
<td>-0.12</td>
<td>-0.11</td>
</tr>
<tr>
<td>Question 5, Task 2</td>
<td>-0.08</td>
<td>-0.08</td>
<td>-0.08</td>
</tr>
<tr>
<td>Question 2, Task 3</td>
<td>-0.24</td>
<td>-0.22</td>
<td>-0.25</td>
</tr>
<tr>
<td>Question 3, Task 3</td>
<td>-0.34</td>
<td>-0.33</td>
<td>-0.34</td>
</tr>
<tr>
<td>Question 5, Task 3</td>
<td>-0.38</td>
<td>-0.40</td>
<td>-0.39</td>
</tr>
</tbody>
</table>

7.18 The Questionnaire

In the following questionnaire, it should be noted that questions 1 - 15 were asked for each task taken by the students. Questions 16 - 19 were separate questions added to the end of the three other questionnaires. This example is from Task 1, with questions 16 - 19 added to the end.
A. Please complete these details.

Name: ________________________________________________________________

Age: _______years, _______months.

Class at School: _______________________________________________________

B. Please complete the following by placing a circle around the most appropriate answer. For example:

Question: It is useful to study the day before an oral test.

Strongly Agree No Disagree Strongly Agree
Agree Opinion

1. I believe that the Picture task would provide an examiner with an accurate idea of my ability to speak English.

Strongly Agree No Disagree Strongly Agree
Agree Opinion

2. I felt nervous before the Picture task.

Strongly Agree No Disagree Strongly Agree
Agree Opinion

3. I felt nervous while I was doing the Picture task.

Strongly Agree No Disagree Strongly Agree
Agree Opinion

4. I believe I did well on the Picture task.

Strongly Agree No Disagree Strongly Agree
Agree Opinion

5. If I had done the Picture task on another day, I would have done better.

Strongly Agree No Disagree Strongly Agree
Agree Opinion
6. I believe that the Picture task provided me with an adequate opportunity to demonstrate my ability to speak English.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>No Opinion</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
</table>

Please explain why: ____________________________________________________________

7. The time allowed for the Picture task was too short.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>No Opinion</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
</table>

8. I liked doing the Picture task.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>No Opinion</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
</table>

Please explain why: ____________________________________________________________

9. I understood what I was supposed to do in the Picture task.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>No Opinion</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
</table>

10. I thought that the Picture task was related to what I learn in class.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>No Opinion</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
</table>

11. If a different teacher had conducted the Picture task, I would have done better.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>No Opinion</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
</table>

12. I thought that the Picture task was too difficult.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>No Opinion</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
</table>
Please explain why: __________________________________________________________

13. I thought that the Picture task was interesting.

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>No Opinion</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

14. I thought that doing the Picture task was an unpleasant experience

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>No Opinion</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

15. Did you think that the subject chosen for the Picture task was particularly fair or unfair? Please give your reasons.

<table>
<thead>
<tr>
<th>Very fair</th>
<th>Fair</th>
<th>No opinion</th>
<th>unfair</th>
<th>Very unfair</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reasons: __________________________________________________________

16. If you were going to take an oral test in an examination, which one of the three tasks would you prefer to do? Put a "1" next to the task you would prefer most, a "2" next to your second choice, and a "3" next to the task you would least like to do.

Task 1: Picture task
Task 2: Discussion of passage
Task 3: Group discussion

17. If you felt nervous during any of the tasks, what would have made you feel less nervous?

______________________________________________________________________

______________________________________________________________________

______________________________________________________________________
18. How would you rate your own proficiency in spoken English?

Generally: Very good  good  average  poor  very poor
Grammatical Accuracy: Very good  good  average  poor  very poor
Fluency: Very good  good  average  poor  very poor

19. For how many years have you been studying English?
Appendix to Chapter Eight

Table 8.1. ANOVA for the condition Fluency

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>SS</th>
<th>DF</th>
<th>MEAN-SQUARE</th>
<th>F-RATIO</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>RATER</td>
<td>72.51</td>
<td>4</td>
<td>18.12</td>
<td>12.90</td>
<td>.001</td>
</tr>
<tr>
<td>TASK</td>
<td>17.34</td>
<td>2</td>
<td>8.67</td>
<td>6.14</td>
<td>.002</td>
</tr>
<tr>
<td>RATER*TASK</td>
<td>15.90</td>
<td>8</td>
<td>1.99</td>
<td>.42</td>
<td>.186</td>
</tr>
<tr>
<td><strong>RESIDUAL</strong></td>
<td><strong>948.22</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Generalizability Coefficients

- Reliability Coefficient: .90
- Inter-rater generalizability coefficient: .93
- Equivalent forms generalizability coefficient: .98
- Forms by Raters generalizability coefficient: .99

Table 8.2. Variation between raters (Fluency)

<table>
<thead>
<tr>
<th>Source (Raters)</th>
<th>SS</th>
<th>DF</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 &amp; 2</td>
<td>0.13</td>
<td>1</td>
<td>0.09</td>
<td>.761</td>
</tr>
<tr>
<td>1 &amp; 3</td>
<td>46.26</td>
<td>1</td>
<td>32.93</td>
<td>.000</td>
</tr>
<tr>
<td>1 &amp; 4</td>
<td>0.44</td>
<td>1</td>
<td>0.31</td>
<td>.577</td>
</tr>
<tr>
<td>1 &amp; 5</td>
<td>11.36</td>
<td>1</td>
<td>8.09</td>
<td>.005</td>
</tr>
<tr>
<td>2 &amp; 3</td>
<td>51.31</td>
<td>1</td>
<td>36.52</td>
<td>.000</td>
</tr>
<tr>
<td>2 &amp; 4</td>
<td>1.05</td>
<td>1</td>
<td>0.75</td>
<td>.388</td>
</tr>
<tr>
<td>2 &amp; 5</td>
<td>13.93</td>
<td>1</td>
<td>9.91</td>
<td>.002</td>
</tr>
<tr>
<td>3 &amp; 4</td>
<td>37.70</td>
<td>1</td>
<td>26.83</td>
<td>.000</td>
</tr>
<tr>
<td>3 &amp; 5</td>
<td>11.77</td>
<td>1</td>
<td>8.38</td>
<td>.004</td>
</tr>
<tr>
<td>4 &amp; 5</td>
<td>7.34</td>
<td>1</td>
<td>5.22</td>
<td>.023</td>
</tr>
</tbody>
</table>

Table 8.3. Variation between tasks (Fluency)

<table>
<thead>
<tr>
<th>Source (Tasks)</th>
<th>SS</th>
<th>DF</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 &amp; 2</td>
<td>1.47</td>
<td>1</td>
<td>1.05</td>
<td>.307</td>
</tr>
<tr>
<td>1 &amp; 3</td>
<td>8.09</td>
<td>1</td>
<td>5.76</td>
<td>.017</td>
</tr>
<tr>
<td>2 &amp; 3</td>
<td>16.45</td>
<td>1</td>
<td>11.71</td>
<td>.001</td>
</tr>
</tbody>
</table>
Table 8.4 ANOVA for the condition Accuracy

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>SUM OF SQUARES</th>
<th>DF</th>
<th>MEAN-SQUARE</th>
<th>F-RATIO</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>RATER</td>
<td>81.97</td>
<td>4</td>
<td>20.49</td>
<td>18.99</td>
<td>.000</td>
</tr>
<tr>
<td>TASK</td>
<td>16.89</td>
<td>2</td>
<td>8.45</td>
<td>7.83</td>
<td>.000</td>
</tr>
<tr>
<td>RATER*TASK</td>
<td>13.15</td>
<td>8</td>
<td>1.64</td>
<td>1.52</td>
<td>.145</td>
</tr>
</tbody>
</table>

RESIDUAL  728.10

Generalizability Coefficients

<table>
<thead>
<tr>
<th>Reliability Coefficient</th>
<th>.87</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-rater generalizability coefficient</td>
<td>.90</td>
</tr>
<tr>
<td>Equivalent forms generalizability coefficient</td>
<td>.98</td>
</tr>
<tr>
<td>Forms by Raters generalizability coefficient</td>
<td>.98</td>
</tr>
</tbody>
</table>

Table 8.5. Variation between raters (Accuracy)

<table>
<thead>
<tr>
<th>Source (Raters)</th>
<th>SS</th>
<th>DF</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 &amp; 2</td>
<td>0.00</td>
<td>1</td>
<td>0.00</td>
<td>.954</td>
</tr>
<tr>
<td>1 &amp; 3</td>
<td>60.29</td>
<td>1</td>
<td>55.90</td>
<td>.000</td>
</tr>
<tr>
<td>1 &amp; 4</td>
<td>3.29</td>
<td>1</td>
<td>3.05</td>
<td>.081</td>
</tr>
<tr>
<td>1 &amp; 5</td>
<td>10.71</td>
<td>1</td>
<td>9.93</td>
<td>.002</td>
</tr>
<tr>
<td>2 &amp; 3</td>
<td>59.36</td>
<td>1</td>
<td>55.03</td>
<td>.000</td>
</tr>
<tr>
<td>2 &amp; 4</td>
<td>3.07</td>
<td>1</td>
<td>2.85</td>
<td>.092</td>
</tr>
<tr>
<td>2 &amp; 5</td>
<td>10.32</td>
<td>1</td>
<td>9.57</td>
<td>.002</td>
</tr>
<tr>
<td>3 &amp; 4</td>
<td>35.75</td>
<td>1</td>
<td>33.14</td>
<td>.000</td>
</tr>
<tr>
<td>3 &amp; 5</td>
<td>19.92</td>
<td>1</td>
<td>18.47</td>
<td>.000</td>
</tr>
<tr>
<td>4 &amp; 5</td>
<td>2.18</td>
<td>1</td>
<td>2.03</td>
<td>.155</td>
</tr>
</tbody>
</table>

Table 8.6. Variation between tasks (Accuracy)

<table>
<thead>
<tr>
<th>Source (Tasks)</th>
<th>SS</th>
<th>DF</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 &amp; 2</td>
<td>1.26</td>
<td>1</td>
<td>1.17</td>
<td>.279</td>
</tr>
<tr>
<td>1 &amp; 3</td>
<td>15.89</td>
<td>1</td>
<td>14.73</td>
<td>.000</td>
</tr>
<tr>
<td>2 &amp; 3</td>
<td>8.19</td>
<td>1</td>
<td>7.59</td>
<td>.006</td>
</tr>
</tbody>
</table>
Table 8.7. ANOVA for the condition Global Assessment

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>SUM OF SQUARES</th>
<th>DF</th>
<th>MEAN-SQUARE</th>
<th>F-RATIO</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>RATER</td>
<td>44.01</td>
<td>4</td>
<td>11.00</td>
<td>8.38</td>
<td>.000</td>
</tr>
<tr>
<td>TASK</td>
<td>32.28</td>
<td>2</td>
<td>16.14</td>
<td>12.30</td>
<td>.000</td>
</tr>
<tr>
<td>RATER*TASK</td>
<td>18.23</td>
<td>8</td>
<td>2.28</td>
<td>1.74</td>
<td>.087</td>
</tr>
<tr>
<td>RESIDUAL</td>
<td>885.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Generalizability Coefficients

Reliability Coefficient  .90
Inter-rater generalizability coefficient  .95
Equivalent forms generalizability coefficient  .97
Forms by Raters generalizability coefficient  .98

Table 8.8. Variation between raters (Global Assessment)

<table>
<thead>
<tr>
<th>Source (Raters)</th>
<th>SS</th>
<th>DF</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 &amp; 2</td>
<td>2.64</td>
<td>1</td>
<td>2.01</td>
<td>.156</td>
</tr>
<tr>
<td>1 &amp; 3</td>
<td>9.06</td>
<td>1</td>
<td>6.90</td>
<td>.009</td>
</tr>
<tr>
<td>1 &amp; 4</td>
<td>9.80</td>
<td>1</td>
<td>7.46</td>
<td>.006</td>
</tr>
<tr>
<td>1 &amp; 5</td>
<td>0.61</td>
<td>1</td>
<td>0.47</td>
<td>.495</td>
</tr>
<tr>
<td>2 &amp; 3</td>
<td>1.92</td>
<td>1</td>
<td>1.46</td>
<td>.227</td>
</tr>
<tr>
<td>2 &amp; 4</td>
<td>22.61</td>
<td>1</td>
<td>17.23</td>
<td>.000</td>
</tr>
<tr>
<td>2 &amp; 5</td>
<td>5.80</td>
<td>1</td>
<td>4.42</td>
<td>.036</td>
</tr>
<tr>
<td>3 &amp; 4</td>
<td>37.70</td>
<td>1</td>
<td>28.72</td>
<td>.000</td>
</tr>
<tr>
<td>3 &amp; 5</td>
<td>14.38</td>
<td>1</td>
<td>10.96</td>
<td>.001</td>
</tr>
<tr>
<td>4 &amp; 5</td>
<td>5.51</td>
<td>1</td>
<td>4.20</td>
<td>.041</td>
</tr>
</tbody>
</table>

Table 8.9. Variation between tasks (Global Assessment)

<table>
<thead>
<tr>
<th>Source (Tasks)</th>
<th>SS</th>
<th>DF</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 &amp; 2</td>
<td>1.83</td>
<td>1</td>
<td>1.39</td>
<td>.238</td>
</tr>
<tr>
<td>1 &amp; 3</td>
<td>29.76</td>
<td>1</td>
<td>22.67</td>
<td>.000</td>
</tr>
<tr>
<td>2 &amp; 3</td>
<td>16.84</td>
<td>1</td>
<td>12.83</td>
<td>.000</td>
</tr>
</tbody>
</table>

Table 8.10. Descriptive Statistics for Fluency

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.42</td>
<td>3.39</td>
<td>4.25</td>
<td>3.46</td>
<td>3.82</td>
</tr>
<tr>
<td>S.D</td>
<td>0.93</td>
<td>1.00</td>
<td>0.55</td>
<td>1.41</td>
<td>1.18</td>
</tr>
<tr>
<td>Range</td>
<td>4.33</td>
<td>4.33</td>
<td>2.33</td>
<td>5.00</td>
<td>4.33</td>
</tr>
</tbody>
</table>
### Table 8.11. Pearson Correlation Matrix (Fluency)

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.68</td>
<td>.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.88</td>
<td>.87</td>
<td>.69</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>.87</td>
<td>.85</td>
<td>.69</td>
<td>.92</td>
</tr>
</tbody>
</table>

### Table 8.12. Spearman Rank Order Correlation Matrix (Fluency)

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.60</td>
<td>.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.87</td>
<td>.77</td>
<td>.60</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>.88</td>
<td>.80</td>
<td>.59</td>
<td>.90</td>
</tr>
</tbody>
</table>

### Table 8.13. Descriptive Statistics for Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.80</td>
<td>2.82</td>
<td>3.75</td>
<td>3.01</td>
<td>3.19</td>
</tr>
<tr>
<td>S.D</td>
<td>0.94</td>
<td>0.93</td>
<td>0.51</td>
<td>1.24</td>
<td>0.89</td>
</tr>
<tr>
<td>Range</td>
<td>4.33</td>
<td>4.33</td>
<td>2.33</td>
<td>4.00</td>
<td>3.33</td>
</tr>
</tbody>
</table>

### Table 8.14. Pearson Correlation Matrix (Accuracy)

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.64</td>
<td>.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.84</td>
<td>.89</td>
<td>.54</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>.82</td>
<td>.86</td>
<td>.55</td>
<td>.86</td>
</tr>
</tbody>
</table>
Table 8.15. Spearman Rank Order Correlation Matrix (Accuracy)

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.58</td>
<td>.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.86</td>
<td>.91</td>
<td>.52</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>.83</td>
<td>.86</td>
<td>.48</td>
<td>.89</td>
</tr>
</tbody>
</table>

Table 8.16. Descriptive Statistics for Global Assessment

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5.49</td>
<td>5.65</td>
<td>5.86</td>
<td>5.11</td>
<td>5.38</td>
</tr>
<tr>
<td>S.D</td>
<td>1.10</td>
<td>1.07</td>
<td>0.65</td>
<td>1.15</td>
<td>1.03</td>
</tr>
<tr>
<td>Range</td>
<td>5.67</td>
<td>5.33</td>
<td>3.00</td>
<td>4.67</td>
<td>4.00</td>
</tr>
</tbody>
</table>

Table 8.17. Pearson Correlation Matrix (Global Assessment)

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.52</td>
<td>.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.87</td>
<td>.79</td>
<td>.51</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>.85</td>
<td>.82</td>
<td>.45</td>
<td>.90</td>
</tr>
</tbody>
</table>

Table 8.18. Spearman Rank Order Correlation Matrix (Global Assessment)

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.33</td>
<td>.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.87</td>
<td>.76</td>
<td>.40</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>.88</td>
<td>.81</td>
<td>.35</td>
<td>.90</td>
</tr>
</tbody>
</table>
Table 8.19. Correlation Coefficients on Fluency: Task 1

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.52</td>
<td>.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.78</td>
<td>.77</td>
<td>.48</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>.74</td>
<td>.79</td>
<td>.52</td>
<td>.85</td>
</tr>
</tbody>
</table>

Table 8.20. Correlation Coefficients on Fluency: Task 2

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.16</td>
<td>.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.76</td>
<td>.76</td>
<td>.10</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>.74</td>
<td>.74</td>
<td>.02</td>
<td>.85</td>
</tr>
</tbody>
</table>

Table 8.21. Correlation Coefficients on Fluency: Task 3

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.47</td>
<td>.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.70</td>
<td>.80</td>
<td>.71</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>.71</td>
<td>.74</td>
<td>.72</td>
<td>.78</td>
</tr>
</tbody>
</table>

Table 8.22. Correlation Coefficients on Accuracy: Task 1

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.39</td>
<td>.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.72</td>
<td>.88</td>
<td>.30</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>.60</td>
<td>.79</td>
<td>.21</td>
<td>.76</td>
</tr>
</tbody>
</table>
Table 8.23. Correlation Coefficients on Accuracy: Task 2

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.82</td>
<td>J</td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.27</td>
<td>.33</td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.67</td>
<td>.75</td>
<td>.28</td>
</tr>
<tr>
<td>Rater 5</td>
<td>.70</td>
<td>.77</td>
<td>.33</td>
</tr>
</tbody>
</table>

Table 8.24. Correlation Coefficients on Accuracy: Task 3

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.91</td>
<td>J</td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.56</td>
<td>.60</td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.78</td>
<td>.80</td>
<td>.46</td>
</tr>
<tr>
<td>Rater 5</td>
<td>.66</td>
<td>.65</td>
<td>.50</td>
</tr>
</tbody>
</table>

Table 8.25. Correlation Coefficients on ELTS Global Assessment: Task 1

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.73</td>
<td>J</td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.49</td>
<td>.35</td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.72</td>
<td>.56</td>
<td>.47</td>
</tr>
<tr>
<td>Rater 5</td>
<td>.73</td>
<td>.65</td>
<td>.53</td>
</tr>
</tbody>
</table>

Table 8.26. Correlation Coefficients on ELTS Global Assessment: Task 2

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.73</td>
<td>J</td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.19</td>
<td>.19</td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.71</td>
<td>.64</td>
<td>.22</td>
</tr>
<tr>
<td>Rater 5</td>
<td>.72</td>
<td>.73</td>
<td>.08</td>
</tr>
</tbody>
</table>
Table 8.27. Correlation Coefficients on ELTS Global Assessment: Task 3

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td>.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>.50</td>
<td>.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>.79</td>
<td>.85</td>
<td>.50</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>.64</td>
<td>.70</td>
<td>.55</td>
<td>.77</td>
</tr>
</tbody>
</table>

Table 8.28. Descriptive Statistics for Fluency, including Rasch severity estimates for raters

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.42</td>
<td>3.39</td>
<td>4.25</td>
<td>3.46</td>
<td>3.82</td>
</tr>
<tr>
<td>S.D</td>
<td>0.93</td>
<td>1.00</td>
<td>0.55</td>
<td>1.41</td>
<td>1.18</td>
</tr>
<tr>
<td>Range</td>
<td>4.33</td>
<td>4.33</td>
<td>2.33</td>
<td>5.00</td>
<td>4.33</td>
</tr>
<tr>
<td>Rater Severity</td>
<td>0.48</td>
<td>0.57</td>
<td>-1.10</td>
<td>0.35</td>
<td>-0.30</td>
</tr>
<tr>
<td>MnSq Outfit</td>
<td>0.8</td>
<td>0.5</td>
<td>1.9</td>
<td>1.0</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Table 8.29. Descriptive Statistics for Accuracy, including Rasch severity estimates for raters

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.80</td>
<td>2.82</td>
<td>3.75</td>
<td>3.01</td>
<td>3.19</td>
</tr>
<tr>
<td>S.D</td>
<td>0.94</td>
<td>0.93</td>
<td>0.51</td>
<td>1.24</td>
<td>0.89</td>
</tr>
<tr>
<td>Range</td>
<td>4.33</td>
<td>4.33</td>
<td>2.33</td>
<td>4.00</td>
<td>3.33</td>
</tr>
<tr>
<td>Rater Severity</td>
<td>0.79</td>
<td>0.77</td>
<td>-1.56</td>
<td>0.22</td>
<td>-0.23</td>
</tr>
<tr>
<td>MnSq Outfit</td>
<td>0.8</td>
<td>0.5</td>
<td>1.9</td>
<td>1.1</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Table 8.30. Descriptive Statistics for ELTS Global Assessment, including Rasch severity estimates for raters

<table>
<thead>
<tr>
<th></th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5.49</td>
<td>5.65</td>
<td>5.86</td>
<td>5.11</td>
<td>5.38</td>
</tr>
<tr>
<td>S.D</td>
<td>1.10</td>
<td>1.07</td>
<td>0.65</td>
<td>1.15</td>
<td>1.03</td>
</tr>
<tr>
<td>Range</td>
<td>5.67</td>
<td>5.33</td>
<td>3.00</td>
<td>4.67</td>
<td>4.00</td>
</tr>
<tr>
<td>Rater Severity</td>
<td>0.04</td>
<td>-0.35</td>
<td>-0.69</td>
<td>0.77</td>
<td>0.23</td>
</tr>
<tr>
<td>MnSq Outfit</td>
<td>1.0</td>
<td>0.8</td>
<td>1.7</td>
<td>0.6</td>
<td>0.7</td>
</tr>
</tbody>
</table>
### Table 8.31. Misfitting ratings

#### (a) Fluency

<table>
<thead>
<tr>
<th>Category</th>
<th>Step</th>
<th>Exp.</th>
<th>Resid.</th>
<th>Rater</th>
<th>Task</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4</td>
<td>4.9</td>
<td>-0.9</td>
<td>3</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>4.8</td>
<td>-1.8</td>
<td>3</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4.0</td>
<td>-2.0</td>
<td>3</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>5.0</td>
<td>-1.0</td>
<td>3</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>4.7</td>
<td>-1.7</td>
<td>3</td>
<td>3</td>
<td>17</td>
</tr>
</tbody>
</table>

#### (b) Accuracy

<table>
<thead>
<tr>
<th>Category</th>
<th>Step</th>
<th>Exp.</th>
<th>Resid.</th>
<th>Rater</th>
<th>Task</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>3.2</td>
<td>-2.2</td>
<td>3</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3.8</td>
<td>-1.8</td>
<td>3</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3.9</td>
<td>-0.9</td>
<td>3</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3.8</td>
<td>-1.8</td>
<td>3</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3.9</td>
<td>-0.9</td>
<td>3</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3.9</td>
<td>-0.9</td>
<td>3</td>
<td>3</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3.6</td>
<td>-1.6</td>
<td>3</td>
<td>3</td>
<td>24</td>
</tr>
</tbody>
</table>

#### (c) ELTS Global scale

<table>
<thead>
<tr>
<th>Category</th>
<th>Step</th>
<th>Exp.</th>
<th>Resid.</th>
<th>Rater</th>
<th>Task</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>3</td>
<td>5.1</td>
<td>-2.1</td>
<td>3</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>2.4</td>
<td>3.6</td>
<td>3</td>
<td>2</td>
<td>32</td>
</tr>
</tbody>
</table>

### Table 8.32. Ability level and likely grades on the Fluency rating scale

<table>
<thead>
<tr>
<th>Logit</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>-3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>-2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>-1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>
Table 8.33. Ability level and likely grades on the Accuracy rating scale

<table>
<thead>
<tr>
<th>Logit</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>-3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>-2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>-1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 8.34. Ability level and likely grades on the ELTS Global rating scale

<table>
<thead>
<tr>
<th>Logit</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>-3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>-2</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>-1</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 8.35. Rasch estimates of judge severity and fit statistics for the Fluency rating scale, taking into account task variability.

<table>
<thead>
<tr>
<th>Rater</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severity</td>
<td>-0.28</td>
<td>0.80</td>
<td>0.51</td>
<td>-0.39</td>
<td>-0.64</td>
</tr>
<tr>
<td>Std. Error</td>
<td>.13</td>
<td>.14</td>
<td>.15</td>
<td>.11</td>
<td>.12</td>
</tr>
<tr>
<td>MnSq Outfit</td>
<td>0.8</td>
<td>0.7</td>
<td>2.4</td>
<td>0.5</td>
<td>0.8</td>
</tr>
</tbody>
</table>
Table 8.36. Rasch estimates of judge severity and fit statistics for the Accuracy rating scale, taking into account task variability.

<table>
<thead>
<tr>
<th>Rater</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severity</td>
<td>0.35</td>
<td>0.25</td>
<td>-0.22</td>
<td>-0.20</td>
<td>-0.18</td>
</tr>
<tr>
<td>Std. Error</td>
<td>.15</td>
<td>.15</td>
<td>.17</td>
<td>.12</td>
<td>.15</td>
</tr>
<tr>
<td>MnSq Outfit</td>
<td>0.9</td>
<td>0.6</td>
<td>1.9</td>
<td>0.6</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 8.37. Rasch estimates of judge severity and fit statistics for the ELTS Global rating scale, taking into account task variability.

<table>
<thead>
<tr>
<th>Rater</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severity</td>
<td>-0.59</td>
<td>0.26</td>
<td>-0.45</td>
<td>-0.47</td>
<td>0.31</td>
</tr>
<tr>
<td>MnSq Outfit</td>
<td>0.8</td>
<td>0.7</td>
<td>2.0</td>
<td>0.7</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Table 8.38. Task difficulty on the Fluency rating scale

<table>
<thead>
<tr>
<th>Task</th>
<th>Difficulty</th>
<th>Std. Error</th>
<th>MnSq Outfit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.36</td>
<td>.10</td>
<td>0.9</td>
</tr>
<tr>
<td>2</td>
<td>0.45</td>
<td>.10</td>
<td>1.2</td>
</tr>
<tr>
<td>3</td>
<td>-0.90</td>
<td>.10</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Table 8.39. Task difficulty on the Accuracy rating scale

<table>
<thead>
<tr>
<th>Task</th>
<th>Difficulty</th>
<th>Std. Error</th>
<th>MnSq Outfit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.13</td>
<td>.11</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>.11</td>
<td>1.0</td>
</tr>
<tr>
<td>3</td>
<td>-0.38</td>
<td>.11</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Table 8.40. Task difficulty on the ELTS Global rating scale

<table>
<thead>
<tr>
<th>Task</th>
<th>Difficulty</th>
<th>Std. Error</th>
<th>MnSq Outfit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.11</td>
<td>.10</td>
<td>0.9</td>
</tr>
<tr>
<td>2</td>
<td>0.22</td>
<td>.09</td>
<td>1.2</td>
</tr>
<tr>
<td>3</td>
<td>-0.11</td>
<td>.11</td>
<td>0.8</td>
</tr>
</tbody>
</table>
### Table 8.41. Intra-Rater correlation coefficients (PPM)

<table>
<thead>
<tr>
<th>Rater</th>
<th>Fluency</th>
<th>Accuracy</th>
<th>ELTS Global scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.75*</td>
<td>.78*</td>
<td>.78*</td>
</tr>
<tr>
<td>2</td>
<td>.91*</td>
<td>.93*</td>
<td>.98*</td>
</tr>
<tr>
<td>3</td>
<td>.53</td>
<td>.18</td>
<td>.47</td>
</tr>
<tr>
<td>4</td>
<td>.97*</td>
<td>.89*</td>
<td>.95*</td>
</tr>
<tr>
<td>5</td>
<td>.89*</td>
<td>.73*</td>
<td>.82*</td>
</tr>
</tbody>
</table>

* Significant at P < .05

### Table 8.42. Intra-Rater correlation coefficients (Spearman)

<table>
<thead>
<tr>
<th>Rater</th>
<th>Fluency</th>
<th>Accuracy</th>
<th>ELTS Global scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.78</td>
<td>.65</td>
<td>.81</td>
</tr>
<tr>
<td>2</td>
<td>.71</td>
<td>.89</td>
<td>.99</td>
</tr>
<tr>
<td>3</td>
<td>.47</td>
<td>.20</td>
<td>.47</td>
</tr>
<tr>
<td>4</td>
<td>.96</td>
<td>.89</td>
<td>.96</td>
</tr>
<tr>
<td>5</td>
<td>.89</td>
<td>.72</td>
<td>.85</td>
</tr>
</tbody>
</table>

### Table 8.43. Intra-Rater descriptive statistics

<table>
<thead>
<tr>
<th>Rater</th>
<th>Fluency</th>
<th>Accuracy</th>
<th>ELTS Global</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Sd.</td>
<td>Range</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>Sd.</td>
<td>Range</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>Sd.</td>
<td>Range</td>
</tr>
<tr>
<td>First</td>
<td>3.75</td>
<td>1.38</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3.38</td>
<td>1.41</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>6.00</td>
<td>1.60</td>
<td>4</td>
</tr>
<tr>
<td>Second</td>
<td>3.75</td>
<td>1.04</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2.88</td>
<td>0.84</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>5.75</td>
<td>1.04</td>
<td>3</td>
</tr>
<tr>
<td>First</td>
<td>3.75</td>
<td>1.49</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>3.38</td>
<td>1.19</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>6.38</td>
<td>1.41</td>
<td>4</td>
</tr>
<tr>
<td>Second</td>
<td>3.63</td>
<td>1.51</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>3.13</td>
<td>1.99</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6.25</td>
<td>1.58</td>
<td>4</td>
</tr>
<tr>
<td>First</td>
<td>4.50</td>
<td>1.31</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>3.75</td>
<td>1.04</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6.75</td>
<td>1.04</td>
<td>3</td>
</tr>
<tr>
<td>Second</td>
<td>4.63</td>
<td>0.52</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>4.00</td>
<td>0.76</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>6.38</td>
<td>0.52</td>
<td>1</td>
</tr>
<tr>
<td>First</td>
<td>4.00</td>
<td>1.51</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>3.50</td>
<td>1.07</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6.13</td>
<td>1.55</td>
<td>4</td>
</tr>
<tr>
<td>Second</td>
<td>3.88</td>
<td>1.55</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>3.50</td>
<td>1.20</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>5.75</td>
<td>1.28</td>
<td>4</td>
</tr>
<tr>
<td>First</td>
<td>4.00</td>
<td>1.31</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>3.13</td>
<td>1.36</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6.00</td>
<td>1.31</td>
<td>4</td>
</tr>
<tr>
<td>Second</td>
<td>4.00</td>
<td>1.60</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>3.63</td>
<td>1.06</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6.00</td>
<td>1.20</td>
<td>4</td>
</tr>
</tbody>
</table>

471
Appendix to Chapter Nine

Table 9.1. Pearson Correlation matrix for the five FCE papers and the oral rating scales.

<table>
<thead>
<tr>
<th>Rating scales</th>
<th>Fluency</th>
<th>Accuracy</th>
<th>ELTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCE papers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>.70</td>
<td>.71</td>
<td>.69</td>
</tr>
<tr>
<td>Writing</td>
<td>.69</td>
<td>.71</td>
<td>.70</td>
</tr>
<tr>
<td>Use of English</td>
<td>.78</td>
<td>.79</td>
<td>.76</td>
</tr>
<tr>
<td>Listening</td>
<td>.55</td>
<td>.59</td>
<td>.53</td>
</tr>
<tr>
<td>Speaking</td>
<td>.76</td>
<td>.76</td>
<td>.75</td>
</tr>
</tbody>
</table>

All correlations significant at p < .01

Table 9.2. Principal Axis Factor Analysis of the FCE papers, Fluency, Accuracy and ELTS.

<table>
<thead>
<tr>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>.79</td>
<td>.66</td>
</tr>
<tr>
<td>Writing</td>
<td>.79</td>
<td>.69</td>
</tr>
<tr>
<td>Use of English</td>
<td>.88</td>
<td>.81</td>
</tr>
<tr>
<td>Listening</td>
<td>.63</td>
<td>.52</td>
</tr>
<tr>
<td>Speaking</td>
<td>.82</td>
<td>.72</td>
</tr>
<tr>
<td>Fluency</td>
<td>.95</td>
<td>.97</td>
</tr>
<tr>
<td>Accuracy</td>
<td>.96</td>
<td>.96</td>
</tr>
<tr>
<td>ELTS</td>
<td>.95</td>
<td>.98</td>
</tr>
</tbody>
</table>

Variance explained by Factors: 5.814
Percentage of total variance explained: 72.67

Table 9.3. Rotated Varimax Factor pattern of FCE papers, Fluency, Accuracy and ELTS rating scales, retaining three Factors.

<table>
<thead>
<tr>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>.41</td>
<td>.12</td>
</tr>
<tr>
<td>Writing</td>
<td>.42</td>
<td>.28</td>
</tr>
<tr>
<td>Use of English</td>
<td>.40</td>
<td>.18</td>
</tr>
<tr>
<td>Listening</td>
<td>.23</td>
<td>1.00</td>
</tr>
<tr>
<td>Speaking</td>
<td>.49</td>
<td>.21</td>
</tr>
<tr>
<td>Fluency</td>
<td>.85</td>
<td>.19</td>
</tr>
<tr>
<td>Accuracy</td>
<td>.82</td>
<td>.23</td>
</tr>
<tr>
<td>ELTS</td>
<td>.88</td>
<td>.17</td>
</tr>
</tbody>
</table>

Variance explained by Factors: 2.95 1.79 2.74
% of total Variance explained: 36.89 22.37 34.25
Table 9.4. Pearson Correlation matrix of the rating components of the FCE speaking test and the Fluency, Accuracy and ELTS rating scales.

<table>
<thead>
<tr>
<th>Rating Scales</th>
<th>Fluency</th>
<th>Accuracy</th>
<th>ELTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCE components</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluency</td>
<td>.67</td>
<td>.63</td>
<td>.64</td>
</tr>
<tr>
<td>Accuracy</td>
<td>.72</td>
<td>.72</td>
<td>.70</td>
</tr>
<tr>
<td>Prosent</td>
<td>.71</td>
<td>.74</td>
<td>.71</td>
</tr>
<tr>
<td>Prosound</td>
<td>.66</td>
<td>.63</td>
<td>.65</td>
</tr>
<tr>
<td>Interact</td>
<td>.74</td>
<td>.74</td>
<td>.73</td>
</tr>
<tr>
<td>Vocab</td>
<td>.67</td>
<td>.69</td>
<td>.66</td>
</tr>
</tbody>
</table>

All correlations significant at $p < .001$

Table 9.5. Principal Axis Factor Analysis of the rating components of the FCE speaking test and the Fluency, Accuracy and ELTS rating scales.

<table>
<thead>
<tr>
<th>Factor 1</th>
<th>$h^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency (FCE)</td>
<td>.81</td>
</tr>
<tr>
<td>Accuracy (FCE)</td>
<td>.88</td>
</tr>
<tr>
<td>Prosent</td>
<td>.89</td>
</tr>
<tr>
<td>Prosound</td>
<td>.82</td>
</tr>
<tr>
<td>Interact</td>
<td>.90</td>
</tr>
<tr>
<td>Vocab</td>
<td>.89</td>
</tr>
<tr>
<td>Fluency</td>
<td>.90</td>
</tr>
<tr>
<td>Accuracy</td>
<td>.90</td>
</tr>
<tr>
<td>ELTS</td>
<td>.89</td>
</tr>
</tbody>
</table>

Variance explained by Factors: 6.908
% of total variance explained: 76.75
Table 9.6. Rotated Varimax Factor pattern of the rating components of the FCE speaking test and the Fluency, Accuracy and ELTS rating scales, retaining two Factors.

<table>
<thead>
<tr>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency (FCE)</td>
<td>.74</td>
</tr>
<tr>
<td>Accuracy (FCE)</td>
<td>.81</td>
</tr>
<tr>
<td>Present</td>
<td>.81</td>
</tr>
<tr>
<td>Prosound</td>
<td>.75</td>
</tr>
<tr>
<td>Interact</td>
<td>.82</td>
</tr>
<tr>
<td>Vocab</td>
<td>.89</td>
</tr>
<tr>
<td>Fluency</td>
<td>.43</td>
</tr>
<tr>
<td>Accuracy</td>
<td>.44</td>
</tr>
<tr>
<td>ELTS</td>
<td>.41</td>
</tr>
</tbody>
</table>

Variance explained by Factors: 4.41, 3.30
% of total variance explained: 48.98, 36.72

Table 9.7. Descriptive statistics for the 7 categories of teacher assessment.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Sd.</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaking</td>
<td>3.74</td>
<td>0.14</td>
<td>3</td>
</tr>
<tr>
<td>Listening</td>
<td>3.89</td>
<td>0.67</td>
<td>3</td>
</tr>
<tr>
<td>Reading</td>
<td>3.54</td>
<td>0.94</td>
<td>3</td>
</tr>
<tr>
<td>Writing</td>
<td>3.24</td>
<td>0.13</td>
<td>3</td>
</tr>
<tr>
<td>Grammatical Accuracy</td>
<td>3.26</td>
<td>1.00</td>
<td>4</td>
</tr>
<tr>
<td>Spelling</td>
<td>3.74</td>
<td>0.80</td>
<td>3</td>
</tr>
<tr>
<td>Homework</td>
<td>3.65</td>
<td>0.92</td>
<td>3</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Rating Scales</th>
<th>Fluency</th>
<th>Accuracy</th>
<th>ELTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher Assessments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaking</td>
<td>.76</td>
<td>.74</td>
<td>.67</td>
</tr>
<tr>
<td>Listening</td>
<td>.70</td>
<td>.66</td>
<td>.73</td>
</tr>
<tr>
<td>Reading</td>
<td>.71</td>
<td>.71</td>
<td>.69</td>
</tr>
<tr>
<td>Writing</td>
<td>.80</td>
<td>.78</td>
<td>.78</td>
</tr>
<tr>
<td>Grammatical Accuracy</td>
<td>.78</td>
<td>.74</td>
<td>.74</td>
</tr>
<tr>
<td>Spelling</td>
<td>.61</td>
<td>.60</td>
<td>.60</td>
</tr>
<tr>
<td>Homework</td>
<td>.75</td>
<td>.71</td>
<td>.72</td>
</tr>
</tbody>
</table>

All correlations significant at p < .001

474
Table 9.9. Principal Axis Factor Analysis of the components of teacher assessment and the Fluency, Accuracy and ELTS rating scales.

<table>
<thead>
<tr>
<th></th>
<th>Factor 1</th>
<th>( h^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaking</td>
<td>.90</td>
<td>.88</td>
</tr>
<tr>
<td>Listening</td>
<td>.78</td>
<td>.70</td>
</tr>
<tr>
<td>Reading</td>
<td>.83</td>
<td>.76</td>
</tr>
<tr>
<td>Writing</td>
<td>.90</td>
<td>.90</td>
</tr>
<tr>
<td>Grammatical Accuracy</td>
<td>.91</td>
<td>.91</td>
</tr>
<tr>
<td>Spelling</td>
<td>.76</td>
<td>.75</td>
</tr>
<tr>
<td>Homework</td>
<td>.89</td>
<td>.87</td>
</tr>
<tr>
<td>Fluency</td>
<td>.93</td>
<td>.93</td>
</tr>
<tr>
<td>Accuracy</td>
<td>.90</td>
<td>.90</td>
</tr>
<tr>
<td>ELTS</td>
<td>.91</td>
<td>.91</td>
</tr>
</tbody>
</table>

Variance explained by Factors: 7.616%
% of total variance explained: 76.16%

Table 9.10. Rotated Varimax factor pattern of the components of teacher assessment and the Fluency, Accuracy and ELTS rating scales.

<table>
<thead>
<tr>
<th></th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaking</td>
<td>.29</td>
<td>.35</td>
<td>.36</td>
<td>.83</td>
</tr>
<tr>
<td>Listening</td>
<td>.36</td>
<td>.40</td>
<td>.15</td>
<td>.60</td>
</tr>
<tr>
<td>Reading</td>
<td>.43</td>
<td>.38</td>
<td>.25</td>
<td>.58</td>
</tr>
<tr>
<td>Writing</td>
<td>.35</td>
<td>.41</td>
<td>.87</td>
<td>.36</td>
</tr>
<tr>
<td>Grammatical Accuracy</td>
<td>.50</td>
<td>.40</td>
<td>.48</td>
<td>.47</td>
</tr>
<tr>
<td>Spelling</td>
<td>.80</td>
<td>.28</td>
<td>.23</td>
<td>.24</td>
</tr>
<tr>
<td>Homework</td>
<td>.71</td>
<td>.36</td>
<td>.24</td>
<td>.47</td>
</tr>
<tr>
<td>Fluency</td>
<td>.31</td>
<td>.84</td>
<td>.26</td>
<td>.35</td>
</tr>
<tr>
<td>Accuracy</td>
<td>.29</td>
<td>.84</td>
<td>.25</td>
<td>.32</td>
</tr>
<tr>
<td>ELTS</td>
<td>.29</td>
<td>.87</td>
<td>.24</td>
<td>.30</td>
</tr>
</tbody>
</table>

Variance explained: 2.17 3.12 1.50 2.32
% of total Variance: 21.71 31.23 14.98 23.20

475
Table 9.11. Statistics for validation by Group Differences on the Fluency scale.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>DF</th>
<th>Mean Square</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>136.24</td>
<td>2</td>
<td>68.12</td>
<td>34.26</td>
<td>.00</td>
</tr>
<tr>
<td>Within Groups</td>
<td>83.50</td>
<td>42</td>
<td>1.99</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Matrix of Absolute Mean Differences</th>
<th>Matrix of Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>Group 2</td>
</tr>
<tr>
<td>Group 2</td>
<td>3.04</td>
</tr>
<tr>
<td>Group 3</td>
<td>4.60</td>
</tr>
</tbody>
</table>

Table 9.12. Statistics for validation by Group Differences on the Accuracy scale.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>DF</th>
<th>Mean Square</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>124.23</td>
<td>2</td>
<td>62.12</td>
<td>26.82</td>
<td>.00</td>
</tr>
<tr>
<td>Within Groups</td>
<td>97.27</td>
<td>42</td>
<td>2.32</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Matrix of Absolute Mean Differences</th>
<th>Matrix of Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>Group 2</td>
</tr>
<tr>
<td>Group 2</td>
<td>2.46</td>
</tr>
<tr>
<td>Group 3</td>
<td>4.37</td>
</tr>
</tbody>
</table>

Table 9.13. Statistics for validation by Group Differences on the ELTS scale.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>DF</th>
<th>Mean Square</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>121.26</td>
<td>2</td>
<td>60.63</td>
<td>27.72</td>
<td>.00</td>
</tr>
<tr>
<td>Within Groups</td>
<td>91.87</td>
<td>42</td>
<td>2.19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Matrix of Absolute Mean Differences</th>
<th>Matrix of Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>Group 2</td>
</tr>
<tr>
<td>Group 2</td>
<td>2.67</td>
</tr>
<tr>
<td>Group 3</td>
<td>4.34</td>
</tr>
</tbody>
</table>
Table 9.14. Summary statistics for the three groups in the validation by Group Differences study.

<table>
<thead>
<tr>
<th>Group</th>
<th>Cases</th>
<th>Fluency Mean</th>
<th>SD</th>
<th>Accuracy Mean</th>
<th>SD</th>
<th>ELTS Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>2.93</td>
<td>1.69</td>
<td>2.82</td>
<td>1.65</td>
<td>2.45</td>
<td>1.93</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>-0.12</td>
<td>1.46</td>
<td>0.35</td>
<td>1.72</td>
<td>-0.23</td>
<td>1.51</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>-1.67</td>
<td>1.19</td>
<td>-1.56</td>
<td>1.21</td>
<td>-1.89</td>
<td>1.14</td>
</tr>
</tbody>
</table>

Table 9.15. MTMM matrix of correlation coefficients

<table>
<thead>
<tr>
<th>Method 1 Pictures</th>
<th>Method 2 Text</th>
<th>Method 2 Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>A1</td>
<td>E1</td>
</tr>
<tr>
<td>Method F1</td>
<td>.94</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>A1</td>
<td>.97</td>
</tr>
<tr>
<td></td>
<td>E1</td>
<td>.97</td>
</tr>
<tr>
<td>Method F2</td>
<td>.92</td>
<td>.92</td>
</tr>
<tr>
<td>2</td>
<td>A2</td>
<td>.89</td>
</tr>
<tr>
<td></td>
<td>E2</td>
<td>.89</td>
</tr>
<tr>
<td>Method F3</td>
<td>.85</td>
<td>.86</td>
</tr>
<tr>
<td>3</td>
<td>A3</td>
<td>.87</td>
</tr>
<tr>
<td></td>
<td>E3</td>
<td>.86</td>
</tr>
</tbody>
</table>

Table 9.16. Squared multiple correlations and coefficients of determination

<table>
<thead>
<tr>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
</tr>
</thead>
<tbody>
<tr>
<td>.97</td>
<td>.98</td>
<td>.94</td>
<td>.97</td>
<td>.94</td>
<td>.95</td>
<td>.96</td>
<td>.98</td>
<td>.95</td>
</tr>
</tbody>
</table>

Coefficient of determination = 1.0

---

477
Table 9.17. Goodness of fit statistics

<table>
<thead>
<tr>
<th>Fit</th>
<th>RMSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>.79</td>
<td>.01</td>
</tr>
<tr>
<td>(X^2): 29.0 ((p = .36))</td>
<td></td>
</tr>
</tbody>
</table>

Table 9.18. \(\Phi\) matrix for the test method facets (\(\xi_1, \xi_2\) and \(\xi_3\))

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.94</td>
<td>1.0</td>
</tr>
<tr>
<td>3</td>
<td>.91</td>
<td>.90</td>
</tr>
</tbody>
</table>

Table 9.19. Factor loadings on \(\xi\) variables for the correlation matrix

<table>
<thead>
<tr>
<th>(\xi_1)</th>
<th>(\xi_2)</th>
<th>(\xi_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>X2</td>
<td>X3</td>
</tr>
<tr>
<td>.32</td>
<td>.03</td>
<td>.01</td>
</tr>
<tr>
<td>.02</td>
<td>.36</td>
<td>.01</td>
</tr>
<tr>
<td>.20</td>
<td>.02</td>
<td>.26</td>
</tr>
</tbody>
</table>

Table 9.20. Modification index for the uncorrected correlation matrix

<table>
<thead>
<tr>
<th>(\xi_1)</th>
<th>(\xi_2)</th>
<th>(\xi_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>X2</td>
<td>X3</td>
</tr>
<tr>
<td>0.00</td>
<td>1.52</td>
<td>0.06</td>
</tr>
<tr>
<td>0.64</td>
<td>0.00</td>
<td>0.11</td>
</tr>
<tr>
<td>0.51</td>
<td>0.16</td>
<td>0.00</td>
</tr>
</tbody>
</table>

478
### Table 9.21. δik statistics for scales and tasks

<table>
<thead>
<tr>
<th>Fluency Band</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>-3.96</td>
<td>-1.72</td>
<td>0.24</td>
<td>1.10</td>
<td>4.34</td>
</tr>
<tr>
<td>Task 2</td>
<td>-3.99</td>
<td>-1.46</td>
<td>0.11</td>
<td>1.73</td>
<td>3.60</td>
</tr>
<tr>
<td>Task 3</td>
<td>-3.88</td>
<td>-1.73</td>
<td>0.07</td>
<td>1.19</td>
<td>3.64</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accuracy Band</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>-4.43</td>
<td>-0.56</td>
<td>1.24</td>
<td>3.76</td>
</tr>
<tr>
<td>Task 2</td>
<td>-3.24</td>
<td>-0.60</td>
<td>0.60</td>
<td>3.25</td>
</tr>
<tr>
<td>Task 3</td>
<td>-3.70</td>
<td>-1.13</td>
<td>1.23</td>
<td>3.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ELTS Band</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>-3.94</td>
<td>-3.45</td>
<td>-1.24</td>
<td>0.43</td>
<td>3.19</td>
<td>5.07</td>
<td></td>
</tr>
<tr>
<td>Task 2</td>
<td>-5.44</td>
<td>-3.56</td>
<td>-1.78</td>
<td>-0.47</td>
<td>1.92</td>
<td>3.70</td>
<td>5.63</td>
</tr>
<tr>
<td>Task 3</td>
<td>-4.18</td>
<td>-4.39</td>
<td>-2.70</td>
<td>-1.11</td>
<td>1.86</td>
<td>3.88</td>
<td>6.65</td>
</tr>
</tbody>
</table>

### Table 9.22. Deviations from the mean δik (δ.k)

<table>
<thead>
<tr>
<th>Fluency Band</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>.02</td>
<td>-.08</td>
<td>.10</td>
<td>-.24</td>
<td>.48</td>
</tr>
<tr>
<td>Task 2</td>
<td>.05</td>
<td>.18</td>
<td>-.03</td>
<td>.39</td>
<td>-.26</td>
</tr>
<tr>
<td>Task 3</td>
<td>-.07</td>
<td>-.09</td>
<td>-.07</td>
<td>-.15</td>
<td>-.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accuracy Band</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>-.64</td>
<td>.20</td>
<td>.22</td>
<td>.22</td>
</tr>
<tr>
<td>Task 2</td>
<td>.55</td>
<td>.16</td>
<td>-.42</td>
<td>-.29</td>
</tr>
<tr>
<td>Task 3</td>
<td>.09</td>
<td>-.37</td>
<td>.21</td>
<td>.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ELTS Band</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>.58</td>
<td>.35</td>
<td>.67</td>
<td>.81</td>
<td>.87</td>
<td>.85</td>
<td></td>
</tr>
<tr>
<td>Task 2</td>
<td>-.92</td>
<td>.24</td>
<td>.13</td>
<td>-.09</td>
<td>-.40</td>
<td>-.52</td>
<td>-.51</td>
</tr>
<tr>
<td>Task 3</td>
<td>.34</td>
<td>-.59</td>
<td>-.79</td>
<td>-.73</td>
<td>-.46</td>
<td>-.34</td>
<td>.51</td>
</tr>
</tbody>
</table>
Table 9.23. Rasch Partial Credit Validity statistics

<table>
<thead>
<tr>
<th>Fluency</th>
<th>Band</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td></td>
<td>.98</td>
<td>.92</td>
<td>.90</td>
<td>.77</td>
<td>.57</td>
</tr>
<tr>
<td>Task 2</td>
<td></td>
<td>.95</td>
<td>.82</td>
<td>.97</td>
<td>.64</td>
<td>.75</td>
</tr>
<tr>
<td>Task 3</td>
<td></td>
<td>.94</td>
<td>.91</td>
<td>.93</td>
<td>.85</td>
<td>.79</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Band</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td></td>
<td>.46</td>
<td>.80</td>
<td>.79</td>
<td>.79</td>
</tr>
<tr>
<td>Task 2</td>
<td></td>
<td>.52</td>
<td>.84</td>
<td>.61</td>
<td>.72</td>
</tr>
<tr>
<td>Task 3</td>
<td></td>
<td>.91</td>
<td>.65</td>
<td>.79</td>
<td>.94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ELTS</th>
<th>Band</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td></td>
<td>.50</td>
<td>.67</td>
<td>.44</td>
<td>.44</td>
<td>.37</td>
<td>.34</td>
<td>.35</td>
</tr>
<tr>
<td>Task 2</td>
<td></td>
<td>.32</td>
<td>.77</td>
<td>.87</td>
<td>.91</td>
<td>.63</td>
<td>.54</td>
<td>.55</td>
</tr>
<tr>
<td>Task 3</td>
<td></td>
<td>.68</td>
<td>.49</td>
<td>.38</td>
<td>.41</td>
<td>.58</td>
<td>.68</td>
<td>.55</td>
</tr>
</tbody>
</table>
Appendix to Chapter Ten

10.1. Interviews with raters

In the following transcripts utterances in [square brackets] are those of the interviewer.

Rater 1

Question 1

[What kind of language do you think each of the three tasks would have elicited?]

Well, the first one with sports was very similar to the kind of things they do in class actually. Although it seems to be easy I don’t think it is, because the first question is "Can you describe the picture?" and the answer you get from the students is "I can see a mountain," "I can see a man climbing." They found it very difficult to describe the pictures, although the students are used to describing pictures like that. However, the language they had to use I believe is... doesn’t require complex language. Yes, I think it is simpler than the other tasks. They can get away with simpler grammar in that kind of exercise; as I said before, they come up with short answers all the time and you have to keep on asking them questions because they just describe what they see in front of them and they expect you to keep on asking questions.

Question 2

[Which of the three tasks did you think was the most difficult/easiest for the students?]
Well the difficult one, I believe, was the second one. Maybe because of the passage. It was a bit long I think first of all. And then the vocabulary in the passage I believe was not the kind of vocabulary you experience from a class unless they happen to have read a passage on that particular topic. The terms used in the passage aren't the kind they come across every day or even in their text books. Yes, and this is the reason that they found it difficult. Of course, I don't know how much time they were given to read the passage. If I remember well, how much time did they have? [Half an hour.] Half an hour. Well, half an hour was more than enough, but then again as I said I believe the terms and vocabulary were a bit difficult for them because they don't read that kind of text. The easiest. The task students seemed to be most interested in was the group discussion, maybe because... the university... tasks like this a a bit difficult. However, they got on very well, maybe because it is something they are interested in, something they have talked about before, they have listened to ideas from other people, they have probably discussed it with their parents, their teachers, and because they were interested they were more fluent. Whether it was easy or not I, well in that respect the way that they found it, they were more fluent. It was easy. Simply because it was within their interests. Then again, the first one, as I said at the beginning, they find it difficult to describe the pictures, but then again when you went on and asked them
to talk about their own sports they found it easier to talk about what they are interested in and what they do themselves, how they employ themselves in their spare time. That particular bit I think is the easiest.

**Question 3**

[Which task produced the most assessable language in terms of (i) quantity and (ii) quality?]

Well, I believe task 3, the discussion on the university, yes, because they had a lot to talk about [So it was the choice of topic?] Yes.

**Question 4**

[Did you think any of the topics were unfair?]

Possibly the second one, for the reasons I gave before.

**Question 5**

[If you could revise any of the three tasks, what would you change to make them better?]

The second one. I would give them a shorter passage and well, I don't know where that passage was taken from anyway. Or the passage could be simplified, and in addition to that the graphs; they found the graphs very confusing possibly because they don't usually deal with things like that. There are some students who follow a course in statistics. In that case they probably find it easy, but then again the graphs themselves were very confusing because of the headings. I think they confused the two graphs and they didn't know what they were talking about. They couldn't read the graphs.
Question 6

[Do you have any other comments on the tasks?]

No, I don’t think so.

Question 7

[Did you find the Fluency scale easy to use?]

Well, can I say something general about the scales? Well, I find them wordy, long, and if you’ve got to go through them while you’re examining a student... I found it difficult anyway, myself. I would rather have something shorter and in a list form. That’s why I have underlined these points because I could easily see something I wanted to. Otherwise, I believe they were detailed enough and they included all the particular points someone should look for when they are examining a student.

Something else I wanted to mention is that you can have something like this with two or three paragraphs in each band and so many papers, yes. For example the first band starts with pauses and when you go on to band 2 you have, let me see, or band 3 the first paragraph talks about misunderstanding a question. The point I’m trying to make here is that the first paragraph in each band should be the first topic. So if pauses is in the first paragraph, it should be in the first paragraph in the second band, and the third band, so that it would be easier for the examiner to look for the points he wants. There should be a kind of sequence to it. That’s all.
Question 8
[Did the Fluency scale contain any descriptors/parts of descriptors which you think are not a part of the concept of fluency?]
I don’t think so.

Question 9
[Did the Fluency scale not contain anything which is, in your opinion, part of the concept of fluency?]
No.

Question 10
[When rating, did you rely on any criteria other than those in the rating scale?]
No, I haven’t thought of anything.

Question 11
[Did you find that you used the entire range of possible bands?]
Yes, yes I did.

Question 12
[Did you think that the amount of detail in the Fluency scale was appropriate?]
No.

Question 13
[omitted for this rater]

Question 14
[Did you find the Accuracy scale easy to use?]
Yes, apart from it being wordy like the fluency scale.
Question 15
[Did the Accuracy scale contain any descriptors/parts of descriptors which you think are not a part of the concept of accuracy?]
No, I don’t think so.

Question 16
[Did the Accuracy scale not contain anything which is, in your opinion, part of the concept of accuracy?]
No.

Question 17
[When rating, did you rely on any criteria other than those in the rating scale?]
No.

Question 18
[Did you find that you used the entire range of possible bands?]
Yes

Question 19
[Did you think that the amount of detail in the Accuracy scale was appropriate?]
Too detailed.

Question 20
[Do you have any other comments on the use of the Accuracy scale or descriptors?]
No, I believe I’ve covered what I wanted to say.
Question 21
[Were there any particular students who you had difficulty rating either in terms of fluency or accuracy? If so, please explain why?]
Yes, I believe it was difficult to place students within the range of 1 to 6. If it was higher then it would be probably easier. 1 to 20 or 1 to 10.

Question 22
[Do you think you are a particularly lenient or harsh rater?] Sometimes harsh, I think it depends. When I find that students keep on staying silent when being asked questions one after the other then I am particularly harsh, and also when they make basic mistakes. In that case again I’m particularly harsh. Lenient? I wouldn’t say I’m lenient. What I would say is I’m probably misled sometimes because of the ability of some students to keep on talking and sometimes they keep on making mistakes, but because they speak so fast sometimes you get the impression that they are fluent even though they are not. So I don’t think I’m lenient in the sense that I feel sorry for them, but I’m misled. At that particular moment I feel they’re good although later on I change my mind.

Question 23
[Do you have any further general comments regarding the process of assessing oral proficiency?]
No.
Rater 2

Question 1
[What kind of language do you think each of the three tasks would have elicited?]
The first one was everyday language requiring everyday vocabulary and expressions about their hobbies and likes and dislikes, and the second and third parts I think they both demanded more intellectual concepts. They had to keep comparing, although the second one was more difficult than the third one because the third one was familiar to them I think.

Question 2
[Which of the three tasks did you think was the most difficult/easiest for the students?]
I think the second one was most difficult and the easiest the first. Most probably, well I thought the second was the most difficult because they could not actually read and read the graph, understand and analyze the graph and the passage was a little bit difficult for them most probably, but... most probably the vocabulary was difficult for them. [So you thought that the picture task was the easiest?] I believe so, because they could all manage to describe what they saw in the picture. Most probably they were used to the activities, although they coped very well with the third one, but I thought they were more comfortable because they were in groups and in the first one they had to talk about their hobbies and
interests which they’ve been doing all along. They were used to that kind of task.

Question 3

[Which task produced the most assessable language in terms of (i) quantity and (ii) quality?]

I think the first one because the second one didn’t, they were not comfortable, they couldn’t actually answer or say what they wanted to say, so there wasn’t very much language for us, so the first one.

Question 4

[Did you think any of the topics were unfair?]

No I didn’t. I thought they were very fair. I think they should be able to handle something like the second task much better than they did.

Question 5

[If you could revise any of the three tasks, what would you change to make them better?]

Most probably I would choose a shorter passage for the second task, and that’s all I would change.

Question 6

[Do you have any other comments on the tasks?]

No.

Question 7

[Did you find the Fluency scale easy to use?]

Yes it was, but I thought of changing the format of the band descriptors. I found that you repeated yourself over and over and I would prefer if we changed the fluency into communication because I believe that communication,
the word communication, shows, covers fluency, and more. Communication is not there in the band descriptors.

**Question 8**

[Did the Fluency scale contain any descriptors/parts of descriptors which you think are not a part of the concept of fluency?]

No, I don't think so.

**Question 9**

[Did the Fluency scale not contain anything which is, in your opinion, part of the concept of fluency?]

No, I can't remember.

**Question 10**

[When rating, did you rely on any criteria other than those in the rating scale?]

No, I didn't; but I was wondering whether we should take pronunciation into account when marking them on fluency, and there was a time when a very good student kept saying a word incorrectly and I didn't know what to do so in the end I disregarded it.

**Question 11**

[Did you find that you used the entire range of possible bands?]

Yes. I found the marking a bit restrictive. Sometimes I found that a student was neither a band 3 nor a 4 and I would have preferred much a greater band range [You mean you would have liked something in between?] In between, yes. I couldn't actually fit them anywhere. It wasn't very often but I found there were cases like that.
Question 12
[Did you think that the amount of detail in the Fluency scale was appropriate?]
They need to be made shorter because in the end I didn’t I couldn’t really remember each point that was written down and I remembered just the most important points that I had to listen for.

Question 13
[Do you have any other comments on the use of the Fluency scale or descriptors?]
I’ve already mentioned the mark range. I would have liked it to be communication, and flexibility. It didn’t take that into account at all. I don’t know whether we could combine it with fluency under communication. But if you keep your mark range I think you should have band descriptors for 0 and 5. Anything else... no, there’s nothing else.

Question 14
[Did you find the Accuracy scale easy to use?]
Yes. We had enough information, something that worried me about the ELTS band which was not clear. The ELTS band was not clear, and I found it very clear and you can watch you can look for each student for each... in each band.

Question 15
[Did the Accuracy scale contain any descriptors/parts of descriptors which you think are not a part of the concept
of accuracy?]
No, I don’t think so.

**Question 16**
[Did the Accuracy scale not contain anything which is, in your opinion, part of the concept of accuracy?]
I don’t think so, no.

**Question 17**
[When rating, did you rely on any criteria other than those in the rating scale?]
No.

**Question 18**
[Did you find that you used the entire range of possible bands?]
Yes, but as I said before I would have preferred more, a wider mark range.

**Question 19**
[Did you think that the amount of detail in the Accuracy scale was appropriate?]
It was appropriate, but most probably it could do with less. I found that it was wordy. You could cut down on the actual explanation, make it more more... shorter. And most probably give the examiner a shorter form, otherwise you have to highlight the main points.

**Question 20**
[Do you have any other comments on the use of the Accuracy scale or descriptors?]
No.
Question 21

[Were there any particular students who you had difficulty rating either in terms of fluency or accuracy? If so, please explain why?]

There was one student who didn’t say very much and I wasn’t sure what mark to give. He was in a group with other students and he didn’t have the chance to say much. I gave him a mark of course. And there were a few more where I thought he was a 4 and then I looked at the band descriptors and I said no, he’s not a 4, then I looked at 3 and I thought no, he’s not a 3 either. And I believe I gave a 3 rather than a 4 in the end.

Question 22

[Do you think you are a particularly lenient or harsh rater?]

I think I was fair, although there was one girl I couldn’t hear at all, although I think she must have been very fluent I gave her a 5 but I wasn’t confident.

Question 23

[Do you have any further general comments regarding the process of assessing oral proficiency?]

No.

Rater 3

Question 1

[What kind of language do you think each of the three tasks would have elicited?]
The first one, with the picture, you mean like present, what tenses what... yes, with the leisure what they do what they enjoy doing what they like doing generally, whereas the second task was trying was more complex obviously. Trying to pick out on a more complex topic something like poverty, and again generally, but it was more complex the poverty one [Do you think there was any difference in the kind of language the students had to use?] Yes [not in terms of complexity] yes, yes because actually they are both using the present tense I don't know exactly how to answer this [that's okay, how about the third task?] The third one which was university wasn't it, that one was difficult, they found it difficult, that one is more, yes that was more complex that was with more "should be" and "if they did this" more conditionals, that was more complex in structures that the first two.

Question 2

[Which of the three tasks did you think was the most difficult/easiest for the students?] The third one was the most difficult discussing the university, and I could see some of them were struggling to keep the discussion going, and the first one was the easiest, the first one with the pictures, and it could be because we often do it with them in class, its topics we have discussed before. [Do you think that the picture format is something they are familiar with as well?] Yes, yes more than passages and...
Question 3
[Which task produced the most assessable language in terms of (i) quantity and (ii) quality?]
I found the first one, I was able to assess them all because they were more fluent, they spoke more, the first one, the second one about poverty I could pick, I was able to distinguish who were the better ones from the not so good ones, and the third one I found a bit difficult to assess the third one, also because they were struggling to keep it going. I found the second one I could distinguish who were the really good ones from the weak ones.

Question 4
[Did you think any of the topics were unfair?]
No.

Question 5
[If you could revise any of the three tasks, what would you change to make them better?]
Well, maybe the third one. Maybe I would have done it a bit different... differently. This was what the students got, and the passage. Well I suppose... I don’t know if I’d be able to... but if I did it, I can’t think off hand, but if I had to change one it would be the third one.

Question 6
[Do you have any other comments on the tasks?]
Question 7
[Did you find the Fluency scale easy to use?] Yes. The fluency was easy to use, and of course I preferred well, the ELTS is more general though. Out of the three ELTS I found is the easiest to use out of the three because it was it was clearer and it was more concise. Yes, whereas here I had to keep reading it over and over. It was difficult to internalise the other two. [You found the ELTS the easiest to use, but did you find it very difficult to use the fluency scale?] No, no, not difficult, it was okay.

Question 8
[Did the Fluency scale contain any descriptors/parts of descriptors which you think are not a part of the concept of fluency?] I didn’t notice anything, no.

Question 9
[Did the Fluency scale not contain anything which is, in your opinion, part of the concept of fluency?] I don’t know, no.

Question 10
[When rating, did you rely on any criteria other than those in the rating scale?] No.

Question 11
[Did you find that you used the entire range of possible
bands?

No, I didn’t use... the bands I used mainly were in the middle, the 3, 4, 5 bands rather than the other two. I didn’t go to the extremes, no.

Question 12

[Did you think that the amount of detail in the Fluency scale was appropriate?]

It’s difficult to be able to... I had to go through and highlight what I thought... it was very difficult to internalise... it was too detailed, whereas points would have been easier.

Question 13

[Do you have any other comments on the use of the Fluency scale or descriptors?]

No. It wasn’t difficult to assess them actually. It was good apart from the fact that it was too long.

Question 14

[Did you find the Accuracy scale easy to use?]

Yes. Actually I found it easier than fluency because I was able to pinpoint when you say with verb forms and grammatical structures, it was easier to pick out the areas.

Question 15

[Did the Accuracy scale contain any descriptors/parts of descriptors which you think are not a part of the concept of accuracy?]

No.
**Question 16**

[Did the Accuracy scale not contain anything which is, in your opinion, part of the concept of accuracy?]

No. Again it was very good descriptors the only thing it was the detail, otherwise it was good once I went through and picked out the main things and highlighted them.

**Question 17**

[When rating, did you rely on any criteria other than those in the rating scale?]

No.

**Question 18**

[Did you find that you used the entire range of possible bands?]

Again, not really. I was towards the middle again. I did use a 2 I remember, but it was 3, 4, 5. Yes. So the top I did use. Not so much the number 1.

**Question 19**

[Did you think that the amount of detail in the Accuracy scale was appropriate?]

It was too detailed.

**Question 20**

[Do you have any other comments on the use of the Accuracy scale or descriptors?]

No.
Question 21

[Were there any particular students who you had difficulty rating either in terms of fluency or accuracy? If so, please explain why?]

There was one pupil, maybe there were two or three that I felt I couldn’t quite because they weren’t quite in the one band but I couldn’t lower them to the lower band judging from some of these, and I think it was the fluency that I had the problem with. [You wanted a half band somewhere?] Yes.

Question 22

[Do you think you are a particularly lenient or harsh rater?]

I’m not harsh, no. I think I tend to be a bit more on the lenient side. [Do you mean to say that you are interpreting the band descriptors in your own particular way?] In a way, yes. Especially when it says that thing about not impeding or if something doesn’t impede understanding or, I sometimes think well it doesn’t. I can still understand and pull them up a bit, whereas maybe it does impede, judging from other teachers when we’ve been doing things together. You have to be more tolerant.

Question 23

[Do you have any further general comments regarding the process of assessing oral proficiency?]
Well, I thought it was generally very good, and all all three, well no the third one might have been difficult but maybe it’s just me or I don’t know. I found that generally it was very good and the descriptors were good but there were too many to keep looking through to mark all at once. But generally I think they were good interviews. It flowed naturally as well, it wasn’t false or artificial like some interviews.

Rater 4

Question 1
[What kind of language do you think each of the three tasks would have elicited?]
Apart from vocabulary, opinions, agreement, disagreement... [Which one would that have been?] I would have thought all three, don’t you think all three would have had that? Depending on the character of the student of course. If he’s willing to talk. Agreement, disagreement, vocabulary.

Question 2
[Which of the three tasks did you think was the most difficult/easiest for the students?]
The one which was most difficult. I thought the second one was. The second task. The first one so - so. They had something to say about that, but the second task I thought was more difficult. And the third they found the easiest I think. Maybe because they knew it was... nearer to what they had been talking about in their general
conversation, and so they knew something about it. [So it was because the topic was more familiar?] Yes.

**Question 3**

[Which task produced the most assessable language in terms of (i) quantity and (ii) quality?]

The third. I think so yes, the third. Again because I think it was generally because they knew about the topic. And it interested them enough to talk and it [inaudible] what they knew through language and vocabulary, so I got more information from the third task.

**Question 4**

[Did you think any of the topics were unfair?]

Unfair. Maybe the second one; but then again it’s because they didn’t know very much about it. Yes. So I suppose it could have been unfair in that respect.

**Question 5**

[If you could revise any of the three tasks, what would you change to make them better?]

That needs thinking. Maybe more guidelines for the second one. All these charts and statistics. It’s just too much for them. More guidelines. More guided questions maybe, to have helped them. Or an explanation more you know apart from the examiner. Because they misunderstood the charts and exactly what they had to talk about. The first one. They had coloured photographs didn’t they? So that wasn’t too bad. Some of the topics were I don’t know, way out. Maybe they didn’t have a lot to talk about. I mean ballooning and the others maybe they didn’t have a lot to
talk about, so if they were topics that they had already you know done, like cycling or swimming or deep sea diving, something like that, maybe they would have had more to say about it.

**Question 6**

[Do you have any other comments on the tasks?]

No.

**Question 7**

[Did you find the Fluency scale easy to use?]

No, not really. No. I think there was too much in each band for me to keep going back and reading them. I found there was too much in each band. I had to read it and re-read it every time you know. To get familiar with it. And then sometimes I wasn’t sure between a two and a three or a three and a four. Very similar. So I had some difficulty there.

**Question 8**

[Did the Fluency scale contain any descriptors/parts of descriptors which you think are not a part of the concept of fluency?]

You mention a lot of grammatical points in fluency. [Can you give me an example?] No, I can’t think of any.

**Question 9**

[Did the Fluency scale not contain anything which is, in your opinion, part of the concept of fluency?]

I made a point here. Some of the students were very good and they understood and they could give back humour you know. And and that wasn’t in it. That’s quite advanced
isn’t it though. Being able to tell a joke and and there was a lot of humour from some of the students. So I don’t know, maybe it could have been put in somewhere. One of the bands. Able to you know interact with the examiner and have this... it’s quite advanced stuff but I don’t know.

**Question 10**

[When rating, did you rely on any criteria other than those in the rating scale?]

No. No. There is this gut feeling of course all the time because being an examiner you expect... but most of it I was just looking at the bands.

**Question 11**

[Did you find that you used the entire range of possible bands?]

Yes. Yes. Yes.

**Question 12**

[omitted for this rater]

**Question 13**

[Do you have any other comments on the use of the Fluency scale or descriptors?]

Apart from being long and I had to keep looking over them again no, no, no, okay.

**Question 14**

[Did you find the Accuracy scale easy to use?]

So - so. Again a bit better than the fluency. There wasn’t very much to look at so it helped. If there was less to look at it would be even better. Would have been
even better. [So you found the amount of detail too much?] Yes. Especially when you’re examining you and you want to be sure of what you’re doing. Yes.

**Question 15**

[Did the Accuracy scale contain any descriptors/parts of descriptors which you think are not a part of the concept of accuracy?]

No. I don’t think so. Can I just remind myself. No, no. Very detailed though. We had to think like did they say much of do and may or say and take. I don’t know that was a bit too...

**Question 16**

[Did the Accuracy scale not contain anything which is, in your opinion, part of the concept of accuracy?]

I think so - pronunciation. There wasn’t anything there and I felt some of the students made mistakes... or were they mistakes... I wasn’t sure because of the pronunciation. And these were Cypriot students. Probably because I know the language I was able to presume what they were saying, but if they were completely different you know from different countries and backgrounds I think it would have been an inhibited area that we could misunderstand. [Would you put pronunciation in an accuracy scale or would you prefer a separate pronunciation scale?] No, I think it would go into accuracy.

**Question 17**

[When rating, did you rely on any criteria other than
those in the rating scale?
No. As I said before there was always this gut feeling, but otherwise I kept looking at the erm... and try to avoid the gut feeling. I don't know if I did it very well just looking at the bands.

Question 18
[Did you find that you used the entire range of possible bands?]
Yes.

Question 19
[Did you think that the amount of detail in the Accuracy scale was appropriate?]
Did you mention verbs? I’ve got down here verbs would have been more noticeable than... Did you mention verbs in the accuracy or was it nouns mostly and tenses. They did make a lot of mistakes with their verbs [Do you think there could have been more detail on verbs?] No. It's just that it you know showed their ability. Sort of stood out you know making mistakes with verbs and tenses.

Question 20
[Do you have any other comments on the use of the Accuracy scale or descriptors?]
No.

Question 21
[Were there any particular students who you had difficulty rating either in terms of fluency or accuracy? If so, please explain why?]
Apart from one girl who was very quiet, but that again I
don’t know in the actual room whether I would have been able to hear her or not, but on the tape she was very very quiet so I couldn’t understand what she was saying for one or two of the tasks. Then there were students who were extremely weak in some tasks but much better in other tasks. So if there was just one task in an examination I wouldn’t have been able to give an accurate mark. So it did show that they were able to talk better in some tasks than other tasks. Maybe because it was more related to what they had been talking about outside, it was fairer. [So do you think that an examination would be better if the students had to do more than one task?] Yes. Yes. Yes. If you had the time you could do two or three. But I don’t think we have the time for that. It’s a matter of time.

**Question 22**

[Do you think you are a particularly lenient or harsh rater?]

Neither lenient nor harsh.

**Question 23**

[Do you have any further general comments regarding the process of assessing oral proficiency?] Apart from the cassette and their being videoed, the students I do think that they were tense. So it was inhibiting for them to have the cassette. Not only for the students I think maybe even for the examiner So I don’t agree with having cassettes. Tasks again. If they were more related to their level and their age it would
have been better. What do you do with speakers. I’ve got here a lot of elaborated vocabulary and able to talk very fluently. We just automatically give them the high mark. But again within their speech they did mark grammatical errors. So that I found that difficult. [Are you suggesting that we should separate lexis and grammar into two separate scales?] I think grammar needs to be broken down. For instance are they using basic grammar for instance the present not the past very often, and not using the third conditional or very advanced grammatical sentences. I think that needs to be broken down. Again just points because while you’re examining it’s difficult to read these points because while you’re examining it’s difficult to read these sentences. But if they were just in points it would be easier to to mark. Oh, while I was marking this, the level. Did you have in mind the level or it was just a general bands for students examining students. I mean could this level be used for GCE your bands. I don’t know it’s... I thought it wasn’t right. Because looking at a student with 5 years of English would have been able to talk on most of the subjects and pretty well, would have got a high mark, not a very high mar, but a high mark to pass. But when it came to the written exam would no way get through a GCE. [Are you suggesting that there should be different band scales for different age groups?] Not age groups but different band scales for different levels of exams. I think so. [Is it not possible to have a band scale which goes from almost

507
zero proficiency to expert speaker, and place all students irrespective of age somewhere along the band scale? It must be very detailed I think, because you would just be giving impression marks. Again, I’m not sure. I mean my impression mark with someone else’s impression mark are completely different. I think it needs better criteria for better levels, rather the level of the student. [Who would you use these band scales with?] This would have been for me FCE or GCE level, but no way no other than that. I wouldn’t use for... I would use different criteria for the lower students and maybe even for the ELTS exam. I would use something else. But not to use this with different levels of students. And having the students in mind I think is very important. If you haven’t taught the level or thinking of the student in you know the level of the student, then again that’s diff... it’s important. I mean for an examiner who hasn’t even taught the level and to use these bands for a specific purpose I think is I think it’s important that the examiner should know the level before examining the student. I don’t know, that’s my opinion.

(Note: In this transcript the rater refers to FCE, which is the UCLES First Certificate in English, and GCE, which is the University of London overseas O Level paper in English Language.)
Rater 5

Question 1

[What kind of language do you think each of the three tasks would have elicited?]

Well, in the first one where they were discussing dangerous sports, I thought that was sort of speculation because you ask why you know why do you think people do this type of sports. And this is why I feel they had some problems because they couldn’t cope with the language I mean the vocabulary and also the functions. The second one is the population and poverty one. I don’t know... The group discussion, well the function was talking about you know education and so on.

Question 2

[Which of the three tasks did you think was the most difficult/easiest for the students?]

Well, at first if I’d just looked at the tasks I would have said that the third one would be the most diff... well you know the second the most difficult and the third one I would have thought was quite difficult and I would have thought that the first one was the easiest. But after seeing the interviews I’ve changed my mind and I think that the third one was the easiest because it was the most relevant to their lives and so on, and their interests, and also they were given quite a lot of ideas and vocabulary, but they were able to use the ideas and
vocabulary. And even quite weak students performed quite well and then I would have said the first task was the next. I wouldn't have said it was easy, but and then the second was the most difficult.

Question 3
[Which task produced the most assessable language in terms of (i) quantity and (ii) quality?]
The third.

Question 4
[Did you think any of the topics were unfair?]
The second one.

Question 5
[If you could revise any of the three tasks, what would you change to make them better?]
The third task I thought was very good. I don't think I would do anything to that one. And probably the second one I think they had problems with the bar graph and and the passage as well. I think possibly I don't know I think I would have changed the topic. I think the same kind of thing could have been given to them with a more familiar topic.

Question 6
[Do you have any other comments on the tasks?]
No, I thought the third one was very good.

Question 7
[Did you find the Fluency scale easy to use?]
No. Well, each band was too long so it was impossible to remember them to memorise them at all, and sometimes in
some ways they didn’t appear to refer to fluency. They didn’t actually refer to the speed of delivery and sometimes you have to be a mind reader in order to decide why someone was pausing. Also, what else. Yes, it’s in band 2 it says that the candidate will "spend less time pausing to plan the grammar of an utterance." Sometimes it it... and then in band 3 it says "pauses will occur when they require thinking time in order to provide a propositionally appropriate utterance." Well I mean you don’t really know why they’re pausing. Probably maybe sometimes they’re at a loss I mean they don’t know what to say, so you can’t really know why somebody pauses.

**Question 8**

[Did the Fluency scale contain any descriptors/parts of descriptors which you think are not a part of the concept of fluency?]

Well, some of it appears to fit more with the vocabulary rather than fluency. You see for instance it says band 5 candidates they demonstrate a confidence in their ability to get things right the first time, but there are weak candidates who look confident and they appear you know they believe that what they’re saying is correct. So it’s and so it’s both.

**Question 9**

[Did the Fluency scale not contain anything which is, in your opinion, part of the concept of fluency?]

Yes. Yes. Well, it mentions pauses but there’s nothing about speed of delivery or ease of delivery, which I
think is mentioned inside the other, in the other bands. Flexibility in ELTS.

**Question 10**

[When rating, did you rely on any criteria other than those in the rating scale?]

In a way, yes. Well, I think I was influenced by the ELTS bands you know, with flexibility and initiative and so on, because this wasn’t really mentioned was it. And the passiveness of the person being interviewed.

**Question 11**

[Did you find that you used the entire range of possible bands?]

No. I found that I hardly used band 1. I did use band 6 occasionally, but bands one and two I didn’t use much.

**Question 12**

[Did you think that the amount of detail in the Fluency scale was appropriate?]

Shorter, in note form.

**Question 13**

[Do you have any other comments on the use of the Fluency scale or descriptors?]

No.

**Question 14**

[Did you find the Accuracy scale easy to use?]

I found that was easier, yes. At least the grammar was easier to use. Yes. Because you know the errors that they make were very clearly set out. Well, basically the grammar descriptors were more concrete and I mean they
were practical and so they were easy to follow, whereas the lexis descriptors were, well, in a way more theoretical.

**Question 15**

[Did the Accuracy scale contain any descriptors/parts of descriptors which you think are not a part of the concept of accuracy?]

No, I was quite happy.

**Question 16**

[Did the Accuracy scale not contain anything which is, in your opinion, part of the concept of accuracy?]

No.

**Question 17**

[When rating, did you rely on any criteria other than those in the rating scale?]

No, I didn’t.

**Question 18**

[Did you find that you used the entire range of possible bands?]

Well, yes I did.

**Question 19**

[Did you think that the amount of detail in the Accuracy scale was appropriate?]

I think again it would have been better in note form, rather than having to read through sentences, to have the detail set out you know in note form in a list it would be better, and then it would be easier to make notes as you are watching the interview.
Question 20

[Do you have any other comments on the use of the Accuracy scale or descriptors?]
No.

Question 21

[Were there any particular students who you had difficulty rating either in terms of fluency or accuracy? If so, please explain why?]
Well, yes. On task 3 the last interview I found it difficult. I don't know if you remember them. They were obviously weak students, but carrying out that task they really did so well on the fluency and accuracy scales. I thought they did really well because of the task. [So you would prefer weaker students to have more difficult tasks?] No, no. I was happy with that because it showed what they were capable of doing which the other two tasks didn't.

Question 22

[Do you think you are a particularly lenient or harsh rater?]
I don't think I'm harsh or lenient. In the middle.

Question 23

[Do you have any further general comments regarding the process of assessing oral proficiency?]
No.
Interviews with students

In the following transcripts utterances in [square brackets] are those of the interviewer.

Student 1

Question 1
[What were you thinking about during the interview?]
First of all I was thinking that I wanted to take advantage of these interviews because it would be it was a good experience for me. It was the first time and er I wouldn't mind I knew I wouldn't be able to have another chance to do a thing before the exams. [When you are actually taking the interview, when you were sitting there, what was going through your mind. What were you concentrating on?] I was trying to guess what the other question would the following question would be. [And when you were speaking, were you thinking about what you were saying or were you thinking ahead all the time?] At times I was thinking ahead. At times I was trying to plan what I was going to say, but er I I was also thinking of what I was saying at the particular moment, the way I was saying generally, like the part when I talked about the operations er and I kind of planned it when you asked the question.

Question 2
[If you could do the interview again, with exactly the
same material, would you do anything differently?]
No. [You were quite happy with it as it is?] Yes.

**Question 3**

[Are you nervous when doing oral tests?]
During the first interview I remember I was a bit nervous, but not during the other ones. [Yes, so it was only the first one] Yes. [Can you imagine circumstances under which you would be very nervous?] Under what circumstances [Yeah] If I was given er for example a picture I couldn’t easily describe because I didn’t know the vocabulary. That would be a difficult situation so, about a matter that I don’t know, that I know nothing about. I think I would be nervous in such a situation. [But you didn’t have any problems with these three, the three tasks which you did?] No, I found them quite easy.

**Student 2**

**Question 1**

[What were you thinking about during the interview?] Whether I was good at it, if I was successful. If I was answering good or if I was bad. [In what way, good?] If I was giving logical answers and if I was giving good expressions... [How do you mean?] Grammar and, basically grammar. Not to make mistakes, stupid mistakes, in grammar. [You mean to say you were trying to think of what you were going to say in advance or you were thinking what you were saying as it came out?] Sometimes in advance. With the picture I had time to think what I
would say in advance, but answering the question it was just what came out. [Anything else?] Well, I tried to find the right words to express something er... sometimes I didn’t find them, like the spot where I said the good English. I can’t find a word to express it. [Yes, so you were looking for words] Yes.

Question 2
[If you could do the interview again, with exactly the same material, would you do anything differently?] Maybe. I don’t think so. Well, now that I’ve looked at it again I know some words that I could use, better words. And if two people do an interview it’s better I think because if you don’t know what to say there comes a time when you don’t speak and you don’t know what to say, the other he speaks and fills the gap. [Did you like in the discussion task having the time to prepare?] Yeah, it was very useful. It was something to think of what we were going to say so we didn’t have a gap. [When you were preparing for that, how did you go about preparing it?] We said that er one was going to er be on one side and the other going to be on the other side; to have a conversation we were going to be on the opposite er... side.

Question 3
[Are you nervous when doing oral tests?] Depends. [On what?] The situation. In an exam I would be nervous because it’s erm you have to do well and you’re nervous if you say something wrong and you lose marks.
Maybe you lose marks so you have to be good. You are under the stress that you have to do well.

**Student 3**

[Well, how well do you think you did?] Well not very well because it was the first time so er, I was a bit worried, anxious and I don’t know. And er as you saw I didn’t er speak much. I mean you had to do all the questioning to make me speak. [So if you could do that interview again you would try to take over the conversation more?] Yes. Yes.

**Question 2**

[If you could do the interview again, with exactly the same material, would you do anything differently?] I don’t know. Maybe bring up some more examples er or experiences from my life er and more ideas.

**Question 1**

[What were you thinking about during the interview?] What a terrible experience (laughter) well er, I don’t know, maybe I was trying to guess the next question to answer er while I was looking at the pictures I was trying to think about the possible questions you know and think about the possible answers erm... I was thinking I think a bit more than er the usual one because I was trying to speak right and not make so many mistakes. And I think that’s why I couldn’t speak very much. [Do you find tasks where you are given planning time easier?] Yes. But it depends on the subject, like the pictures I
don’t think I need to prepare myself. I mean they were easy questions about my life so I didn’t have to be prepared, but about the other subjects I think it was easier because it was about the university so I had to plan what I was going to say. To organise my ideas and...

**Question 3**

[Are you nervous when doing oral tests?]

I don’t know. Maybe because it was the first time, the first experience about er that sort of things. I mean er I have never had an interview before er. I guess that’s all. [Can you imagine yourself doing an interview where you felt nervous for some other reason?] No, no.

**Student 4**

**Question 1**

[What were you thinking about during the interview?] er I was thinking what I was going to say. I was rather nervous. I wasn’t sure about what I was saying, because I wasn’t interested in some things and I kept changing my mind. [So you were thinking mainly about ideas?] Yes, and how I was going to express them. [When you say "how" you were going to express them, what do you mean?] Phrases, grammar, some vocabulary I was going to use. These things.

**Question 2**

[If you could do the interview again, with exactly the same material, would you do anything differently?]

Yes, I want to do something different but I don’t think I
would do anything different. Maybe I would talk louder. I don't think I would do anything else. [Yes, you have got a very quite voice, and are you shy generally?] Yes. [Do you think that would influence an examiner?] Yes, definitely, because if I don't have fluency that means I won't get any marks. [So just for an interview you think that you have to alter your character?] Yes. I try to but I don't think I can do anything different, because I don't feel it's natural. When I talk with others I feel it's more natural than in the interview. I just don't like it. [I hope you don't mind if I follow this one. When you're outside say just with friends, you don't have the same problems with talking and expressing yourself?] It depends with what friends because let's say er the friends here at the English Institute are not too close to me but the others at school are. I've known them for many years so it's different. But still I'm shy. In the interview because I'm shy I'm nervous because I think that I won't like it. That's why. I'm nervous and I don't talk. [Can you think of anything which the interviewer could look for when you go for an interview that would tell him or her that you were shy rather than unable to talk freely? Do you understand what I mean?] Yes. Well, it depends on the examiner. If he or she is just looking for er grammar or vocabulary erm they won't understand that I'm shy. But if he knows something about psychology they would understand.

Question 3
Student 5

At first when I finished the interview I thought I had talked a lot, that I had said some nice things. But now I think I could say a lot more, or better. It was a bit funny because I didn’t think I was talking like that. [Oh, you mean the sound of your own voice?] Yes. [Well, unless you record yourself many times it’s difficult to imagine what you sound like.] Yes.

Question 1

[What were you thinking about during the interview?] Well, erm, I was thinking about the pictures er the questions, but I didn’t have er the time to realise what I had to say, er I had to answer immediately so I didn’t think a lot before er I answered the questions. I think if I had more time to think it would be better. It would take more than half an hour then. [When you were actually talking what were you thinking about?] I was trying to think ahead. I just made a lot of mistakes while I was talking because I was trying to think what I was going to answer to the next question. I didn’t know the next question but I tried to guess.

Question 2

[If you could do the interview again, with exactly the same material, would you do anything differently?]
Yes, I would say erm more or less the same things, but I would express them in another way. A better way. [In the task where you had time to prepare, did that help you?] Well, not much because in that ten minutes I had to read the passage that I was given and I had to try to understand what it was talking about. But it depends. It was the the subject was quite easy and familiar so I didn’t have a problem with that.

Question 3

[Are you nervous when doing oral tests?] No, not during it. Before it and afterwards. But not during it. [Why?] I don’t know. Before it’s because I don’t know what it will be like, the questions I have to answer erm, and afterwards because I try to think if I er answer it right or wrong and what I should have said and what not, but not during it. [Is there anything anyone could do to make you feel less nervous?] No. [No, so it’s just one of those things, a fact of life that you have to live with?] Yes.

Student 6

Question 1

[What were you thinking about during the interview?] I think I was trying to concentrating er but there were other thoughts in my mind too. School, future er the exam. [So you weren’t thinking about what you were saying at that particular moment?] No, I didn’t think it was important to do that but I tried to do well. [Did you
like having time to prepare on some tasks? Did it make things easier for you?] I don’t think so, I don’t think too must about tests.

Question 2
[If you could do the interview again, with exactly the same material, would you do anything differently?] Find out some useful vocabulary and erm expressions and probably more fluent erm better pronunciation.

Question 3
[Are you nervous when doing oral tests?] A bit. It was the first time I had such a test. So I was uneasy. [So it will go away with practice.] No. I am a foreigner so I don’t speak English every day. It’s hard to think in English. [So, it doesn’t matter what anybody does, you would always feel nervous, or could it be taken away?] Yes, I think sometimes the atmosphere in the exam is important. I mean at the interview in the class we do with our teacher I don’t feel nervous at all.

Student 7

Question 1
[What were you thinking about during the interview?] When it was going to end. The camera. It was the first time to have a camera and erm... it was stressful.

Question 3
[Are you nervous when doing oral tests?]
Also I was talking in English and this is not usual for me. I don't do it very often and it was like having a new experience. [Is there anything an examiner could do to make you feel more relaxed?] Well erm having a coke in the exams. A chewing gum. That would make things better, and just to er be friends with the interviewer because if you know him or have met him before then you feel more relaxed.

**Question 2**

[If you could do the interview again, with exactly the same material, would you do anything differently?] Well, I would know what to say. [No, assume you were doing it again for the first time.] Oh, well if I wasn't so embarrassed I would say something else. Try to say something more interesting. Just this. You see the subject of sports is very common and each one say erm not very wide there isn't a wide range to select to talk about like different kinds of sport [So you were happier with the other topics?] Not exactly happy. If I had a better subject like computers that I know that I'm interested in I would say more.

**Student 8**

I think I had a lot of problems with the picture. I made a lot of grammatical mistakes there and actually er it's a lack of vocabulary, mostly. In a way, I mean when I said it's difficult to climb I mean the way the mountain was. It was too vertical and he couldn't climb it easily.
That's what I wanted to say but I didn't know the words. [The word didn’t come at that moment] Yes [You mean you didn’t know the word at all or it just didn’t come at that moment?] Sometimes I forget a word. Lots of words I don’t know even simple words I forget them. When I talk, because when you talk it’s not like when you’re writing. When you’re writing you have time to think and if you write it down you read it again while when you talk you don’t have the time to think and erm...

Question 1
[What were you thinking about during the interview?] Well, I didn’t think of anything special, just the interview.

Question 2
[If you could do the interview again, with exactly the same material, would you do anything differently?] erm I would try to describe the picture better. I don’t think I described it very well, because I talked about the hills and the background and I didn’t know how to describe the man because my brain stopped. It was frozen and I didn’t know what to say. Next time I’ll be more careful. [Do you prefer to have time to prepare as in the discussion on the university?] Yes. If you had time to prepare it would be all right, yes. Time to make up your mind I think it’s okay.

Question 3
[Are you nervous when doing oral tests?] Well, it could be an easy thing, but it usually I can’t
talk I can't pronounce words correctly because my tongue gets a bit... and er and also when you talk you have to think fast, you have to talk a lot, talk fast, and if it isn't your native language, your first language, er you have to think a bit more. It's difficult to do all these things at the same time and talk er good English. [So the anxiety you felt is something that no one can do anything about. It's there because you are learning a second language?] I think if the er the examiner is friendly and er an he doesn't keep marking you all the time you'll feel like you're talking to a friend of yours or someone and you won't be that nervous and you won't think about the marks. It's a conversation and not a marking session so I think you'll do much better. If a person is friendly I think it's okay. If he drinks tea all the time or talks with an Oxford accent then...

Table 10.3. Ability level of the sample in logits, and sex.

<table>
<thead>
<tr>
<th>Student</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>4.32</td>
<td>1.90</td>
<td>3.02</td>
<td>N/A</td>
<td>1.64</td>
<td>1.64</td>
<td>1.16</td>
<td>4.68</td>
</tr>
<tr>
<td>Accuracy</td>
<td>4.16</td>
<td>1.58</td>
<td>3.20</td>
<td>N/A</td>
<td>1.74</td>
<td>1.89</td>
<td>0.99</td>
<td>3.88</td>
</tr>
<tr>
<td>ELTS</td>
<td>3.62</td>
<td>0.82</td>
<td>2.46</td>
<td>N/A</td>
<td>1.30</td>
<td>0.82</td>
<td>0.67</td>
<td>3.79</td>
</tr>
<tr>
<td>Sex</td>
<td>M</td>
<td>M</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
</tbody>
</table>

526